University of Bath

PHD

Molecular modelling of antibody combining sites

Whitelegg, Nicholas R. J.

Award date:
1998

Awarding institution:
University of Bath

Link to publication

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

Copyright of this thesis rests with the author. Access is subject to the above licence, if given. If no licence is specified above, original content in this thesis is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC-ND 4.0) Licence (https://creativecommons.org/licenses/by-nc-nd/4.0/). Any third-party copyright material present remains the property of its respective owner(s) and is licensed under its existing terms.

Take down policy
If you consider content within Bath's Research Portal to be in breach of UK law, please contact: openaccess@bath.ac.uk with the details. Your claim will be investigated and, where appropriate, the item will be removed from public view as soon as possible.

Download date: 06. May. 2023

Molecular Modelling of Antibody

Combining Sites

Submitted by Nicholas RJ Whitelegg
for the degree of
Doctor of Philosophy
University of Bath

September 17, 1998

COPYRIGHT

Attention is drawn to the fact that copyright of this thesis rests with its author. This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and no information derived from it may be published without the prior written consent of the author. This thesis may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purpose of consultation.

UMI Number: U115929

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI U115929
Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

UNIVERSITY OF BATH
LIBRARY
56 25 NOV 1993

CONTENTS

Abstract

Antibodies are an important part of the body's defence against invading organisms, and therefore have a number of important therapeutic and research uses. To effectively engineer or humanise antibodies, it is important to know their structure, and the technique of computer modelling can be used to achieve this aim. Modelling has an advantage in terms of speed over the alternative, X-ray crystallography, although at present it is not always as accurate.

The structure of the variable region of antibodies consists of a relatively conserved framework, and the hypervariable loops or complementarity determining regions (CDRs). The fact that the framework is relatively conserved, and that five of the six CDRs often fall into a small set of conformations (canonical structures) was used by Martin, Cheetham and Rees (1989) in their antibody modelling algorithm, which has been released as the package AbM. The remaining CDR, H3, poses the greatest problem as canonical classes have not yet been worked out. Martin, Cheetham and Rees tackled its modelling using a combined database/conformational search (CAMAL).

However, its success was rather mixed, with the modelled conformations not always the closest to the crystal structure. The work here sought to improve the modelling procedure by improving existing algorithms and, where necessary, developing new methods. Analysis was carried out on all the terms contributing to the energies of CDR loops to identify sources of high energy in otherwise 'good' loops, and alternatives to the Martin, Cheetham and Rees solvent-modified potential were explored to model the solvent effect, including the methods of Eisenberg and McLachlan and the Honig group.

The energy analysis identified two principal problems: high energy 1,4-interactions, and energetic clashes between H3 sidechains. In view of these observations, a modified modelling algorithm has been developed in which the 1,4-interactions are ignored in the energy screen and the H3 sidechains built only after minimisation of the backbone. Two methods were employed to build the sidechains; the existing CONGEN iterative method and the dead-end method of Lasters and co-workers, a complex series of algorithms implemented by the author using the published papers of Lasters only. A selection screen, based either on lowest energy or a residue accessibility profile closest to known H3 loops of that length, was developed. In addition, a canonical feature of

H3 loops, not previously described, was discovered and used in the screening process.

Both sidechain build methods performed similarly, modelling aromatic residues best and small and charged residues worst. Of the selection screens, the accessibility profile performed best at selecting models closest to the crystal structure. Indeed, for loops of 10 residues of less, it was particularly effective at rejecting the inaccurate models. The canonical feature screen was valuable for one antibody whose H3 conformation was intractable by the other methods.

In summary, therefore, the modified antibody modelling algorithm presented here performs considerably better at selecting conformations close to the crystal structure in almost all cases than the original AbM algorithm.

Acknowledgements

First, I would like to thank Professor Tony Rees for accepting me on the PhD at quite short notice, and for three years of good and helpful supervision. I would also like to thank Andrew Henry for help with UNIX in the first year, and Paul Calleja for help with general protein modelling questions.

I would also like to particularly thank David Osguthorpe for a good deal of help with the use of the VFF program in loop energy minimisation, and in the area of energetics in general. I would also like to thank Pnina Dauber-Osguthorpe for some useful suggestions in the area of loop energetics.

Andrew Martin, one of the original developers of the combined antibody modelling algorithm, deserves a special thank you for guiding me through the workings of the various AbM files, and for supplying a range of very useful file utilities, as well as several helpful e-mail discussions, as does Steve Searle, who has also worked on developing the algorithm, and has been a great deal of help in discussions on CDR-H3 modelling and the reasons why the modelling may not be as accurate as it should be! I

would also like to thank Andrew and Steve for use of their their thesis figures (Figure 2, and Figures 1 and 8 respectively).

The funding was provided by a University Bursary, and I would like to thank the University for providing this. Thanks also to my supervisor Tony Rees for providing funding for the Asilomar CASP-2 conference which I attended, and found useful for insight into general protein modelling.

Finally I would like to thank my family and friends for providing support during these three years.

Abbreviations

RMSD	Root mean square deviation (Global backbone root-mean-square deviation between CDR-H3 conformations unless otherwise shown)
CDR	Complementarity-determining region
Å	Angstrom
AbM	Antibody Modeller
VFF	Valence Force Field
CAMAL	Combined antibody modelling algorithm
PDB	Protein Data Bank (Bernstein et al, 1977)
N	Nitrogen
H	Hydrogen
CA	C-alpha
CB	C-beta
C	Carbon
O	Oxygen

List of figures

List of tables

CHAPTER 1 - INTRODUCTION

1.1 The immune response

The immune response is the mechanism by which the body destroys foreign organisms. It consists of two stages: an initial recognition stage, usually where protein or carbohydrate, or occasionally nucleic acid, *antigens* on the surface of the foreign body are recognised, and an effector stage where the foreign body is destroyed.

Types of immunity

There are two types of immunity, *innate* and *specific*. *Innate immunity* is effected by molecules and cells always present in the body, and non-specifically destroys foreign organisms, whereas *specific immunity* is effected by cells and/or molecules specific for a particular foreign organism (by way of a surface antigen) and which increase in concentration in response to the organism. The cells involved are (frequently) cytotoxic T cells while the molecules involved are known as immunoglobulins or antibodies.

Innate immunity

The innate component to the immune response is the first to be mobilised, before the specific component can be activated. Cells that surround and digest foreign bodies, phagocytes and macrophages, are involved, as well as natural killer cells, which cause cell destruction by creating membrane pores. Another component of innate (as well as specific) immunity is the complement system (Figure 1) . This is a group of proteins forming a cascade proteolytic pathway, culminating in the formation of the *Membrane Attack Complex* protein, which destroys foreign organisms by creating holes in cell membranes. Complement is activated by two pathways, the classical and alternative pathways, which are involved in the innate and specific responses respectively. The mechanism of activation of the latter will be discussed below, but the former is activated by antigens on the foreign cell surface.

A further function of complement in innate immunity is a process known as *opsonisation*, whereby the invading cell is marked by complement for more effective destruction by phagocytes and macrophages.

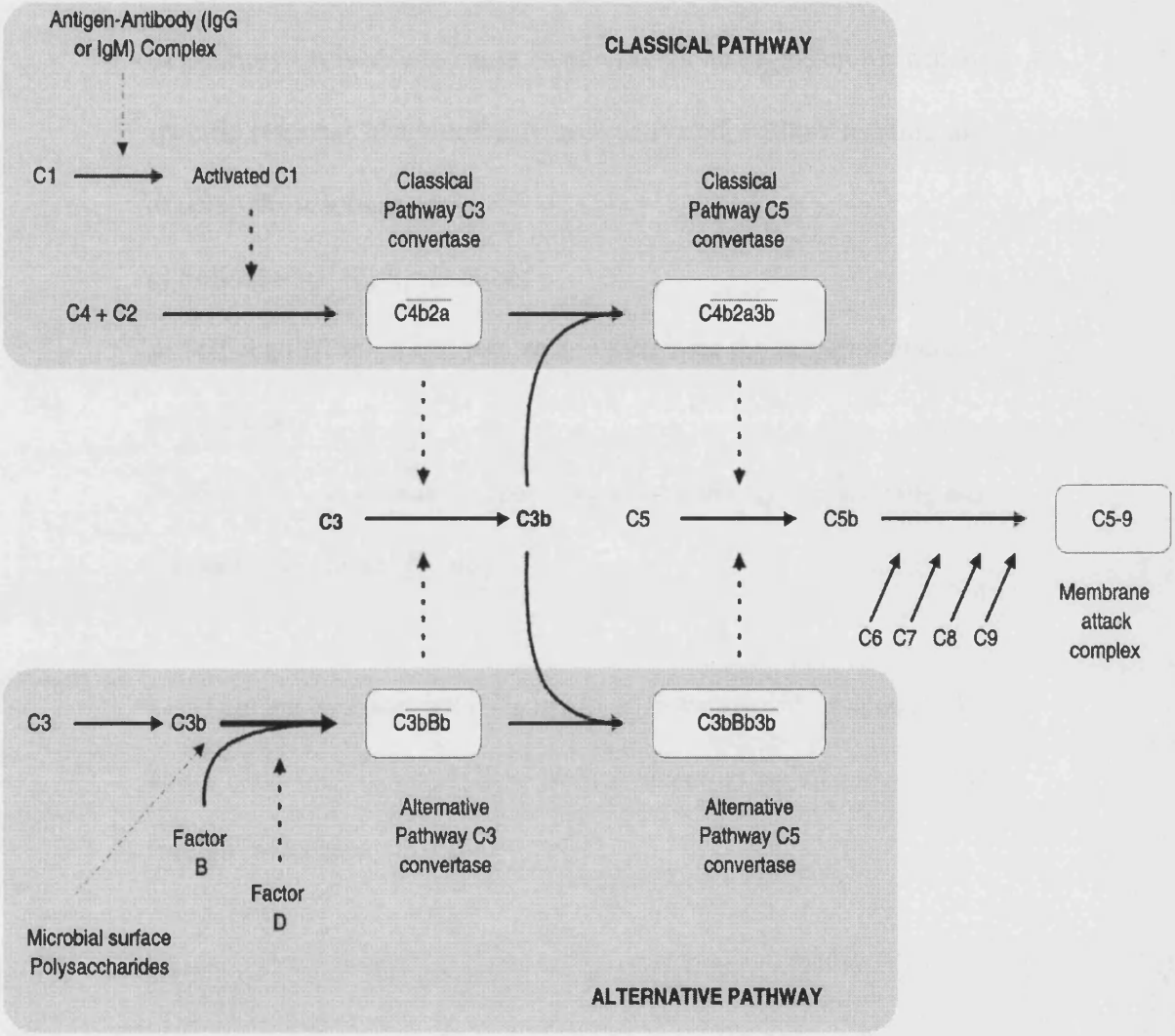


Figure 1. The classical and alternative complement pathways.

Specific immunity

Specific immunity has a number of characteristics which make it an effective means of removing infective organisms:

a) Specificity - it can identify a unique antigen;

b) Memory - it is able to cause destruction of an organism for which a specific response had previously been activated, without needing to develop the response again;

c) Self/non-self discrimination;

d) Self-regulation - it can initiate and terminate the response at the correct time;

e) Diversity - an immune response can be raised against virtually any biological or chemical entity.

There are two types of cells involved in the specific response, B and T cells, both of which have surface receptors which can recognise foreign molecules.

i) B cells

B cells have two roles. Firstly, they act as *antigen presenting cells (APCs)*. The antigen binds to antibody molecules on the surface, is internalised and digested, and fragments presented, in association with major histocompatibility complex (MHC; see below) molecules, to a

sub-type of T cells (see below) known as T helper cells. These act as helpers in proliferation and differentiation of B cells in order to serve their second function: production of antibodies. *Plasma B cells* produce antibodies immediately in order to combat the infection, whereas *memory B cells* remain in order to effect a fast specific immune response in the case of future infection.

Both membrane bound and soluble antibodies are produced by the B cells. Membrane bound antibodies act in antigen presentation, as discussed above, whereas soluble antibodies bind to antigens on foreign cells in solution, causing their destruction in two ways. Firstly the classical complement pathway can be activated, by binding of the antibody to the first complement component; and secondly, opsonisation can take place, in which the antibody (with foreign cell bound) binds to a phagocyte, leading to digestion of the foreign cell.

ii) T cells

There are three subdivisions of T cells, with different functions: *cytotoxic T cells*, which bind to virally infected cells, for example, causing their destruction; *T helper cells* which bind to antigen presenting cells such as B cells and enhance B cell proliferation and antibody production (see above) by lymphokine secretion; and *T suppressor cells* which suppress the immune response to a particular antigen.

T helper and suppressor cells, as well as recognising a specific antigen, also need to recognise the cell type as an APC. A set of polymorphic molecules known as the major histocompatibility complex (MHC) antigens, involved in antigen presentation, fulfil this role (specifically a subtype known as MHC-II). The APC internalises and digests the antigen, a fragment of which then binds to the MHC-II specific for it, and finally the MHC-II/antigen complex is externalised on the APC surface, leading to T helper cell binding. Another subtype, MHC-I, is involved in the action of cytotoxic T cells. Viral proteins are digested within the infected cell, the fragments bind to MHC-I and the complex is externalised on the APC surface. The cytotoxic T cell then binds and kills the infected cell, both by *perforin* secretion which forms pores in the infected cell, and *tumour necrosis factor* secretion which causes programmed cell termination. MHC-I has a further role, in the recognition of cells as self. Each individual has a specific MHC-I molecule expressed on the surface of all body cells, and any cells with a different MHC-I will be destroyed. This is known as the *allotypic* response, and is the deleterious process which takes place when transplants are rejected.

Types of antibody

The antibodies produced in the specific response to infection are known as immunoglobulins M (IgM) and G (IgG). There are a

number of other antibody classes with different functions; IgE is produced in the specific response to allergens, such as pollen, and plays a part in histamine release; IgM is the receptor for antigen binding on B cells and is also produced as soluble antibody in the primary response; IgA is the primary line of defence against organisms, found in secretions such as saliva and mucus; and IgD has a regulatory role in B cell differentiation.

In summary, the co-operation between the different components of the immune system leads to a highly effective means of combatting infection.

1.2 Antibody structure

General antibody (IgG) structure

The IgG molecule is a symmetrical Y shaped molecule (Figures 2 and 3), consisting of two heavy and two light chains. There are two types of light chain, kappa (κ) and lambda (λ), which have different genetic coding. The 'arms' of the Y are known as the *Fab* portion, each arm consisting of a light chain paired with the topmost part of the corresponding heavy chain, whereas the 'tail' of the Y is referred to as the *Fc* portion, consisting of the lowermost part of the heavy

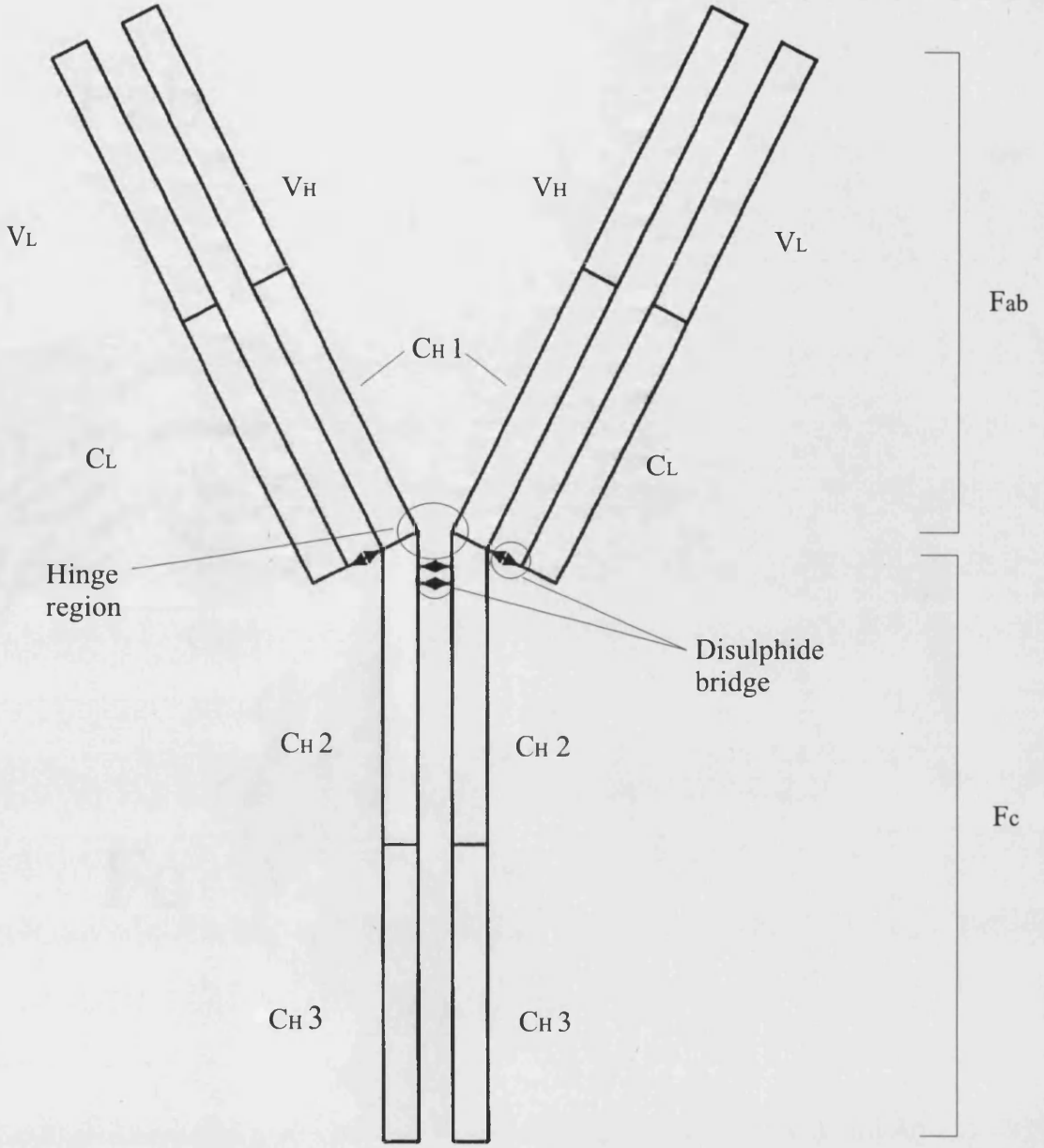


Figure 2. The 2D structure of an antibody, showing the domains.

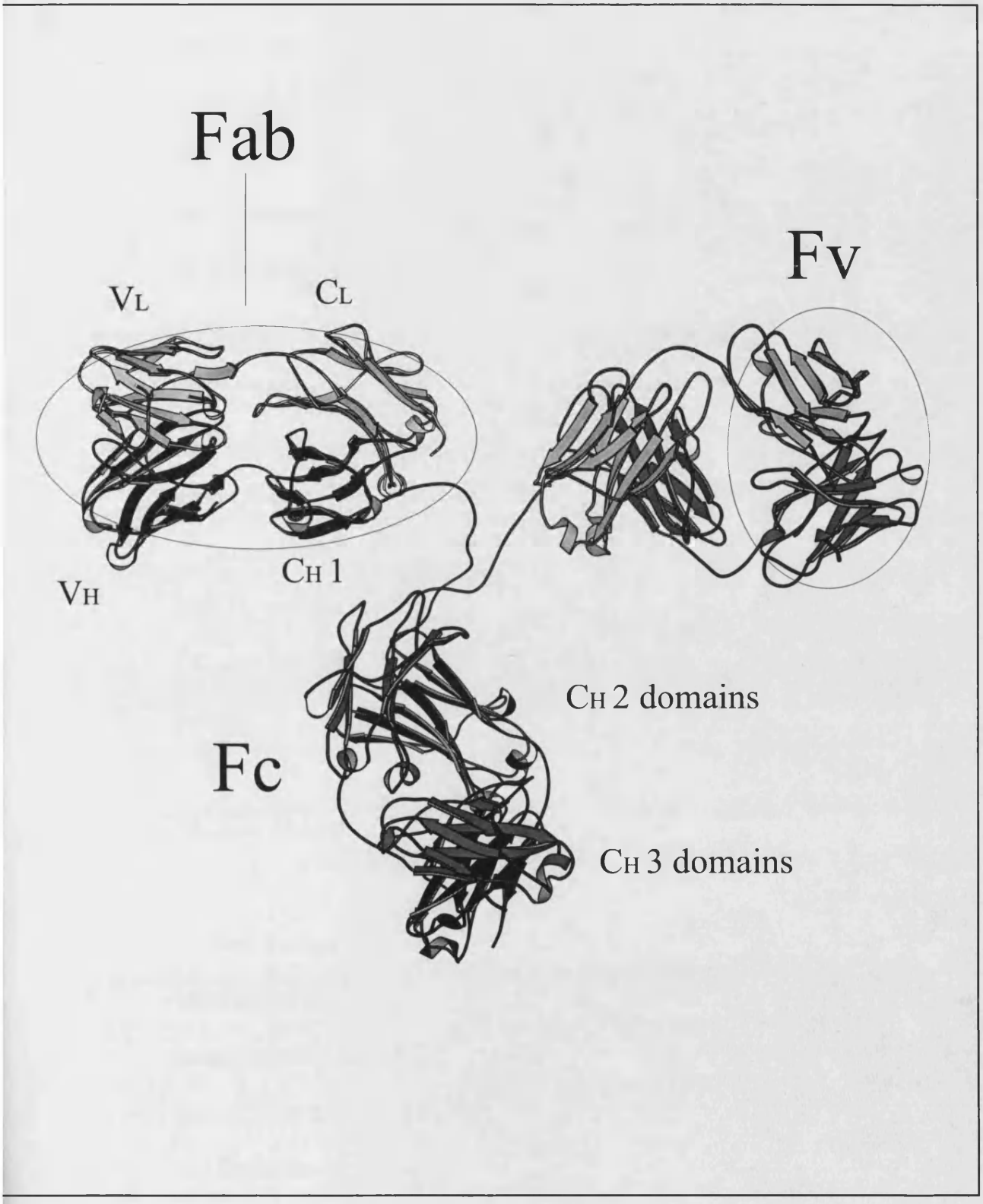


Figure 3. The 3D structure of an intact antibody present in the Protein Data Bank (1IGT), showing the domains and the Fab, Fv and Fc regions.

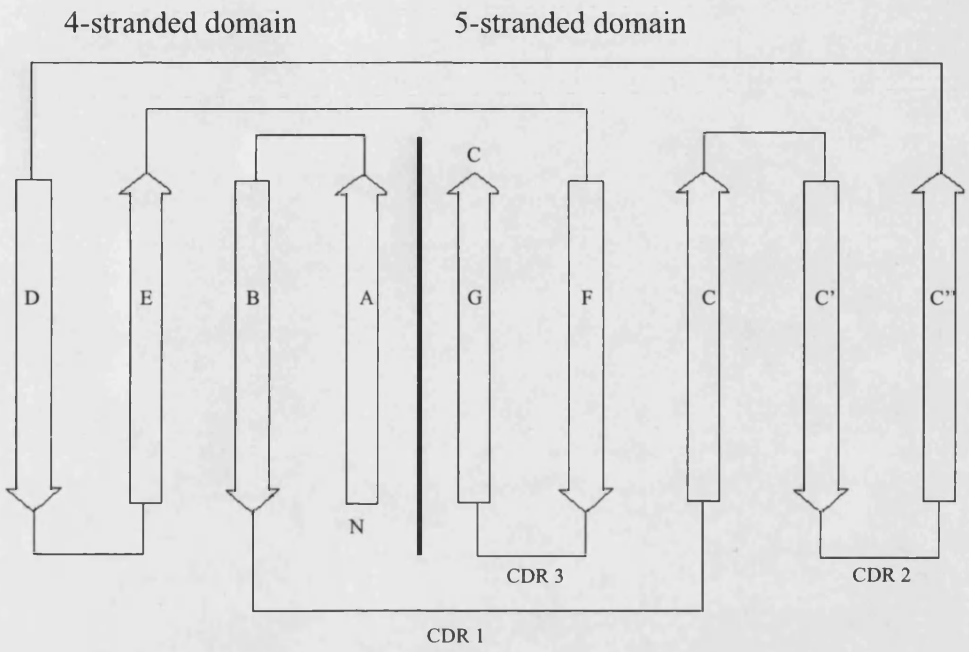
chains, paired with one another. In between the Fab and Fc is a flexible, *hinge region.*

Each heavy chain consists of four domains (V_H and C_H1 in the Fab, and C_H2 and C_H3 in the Fc), whereas each light chain consists of two domains (V_L and C_L). The V_H and V_L domains, at the 'tips' of the Fab, form the *variable region (Fv)*; this is much more variable in sequence than the remainder of the molecule, and is the region which binds antigen. The Fc binds to components of the immune system which destroy invading organisms, such as phagocytes, and complement molecules.

Each domain interacts with a corresponding domain; for example each V_H interacts with the corresponding V_L, each C_H1 interacts with a C_L, and the C_H2 and C_H3 domains each interact with an identical partner.

Structure of the Fv in detail

Each domain (the V_H and V_L) consists of a common structural motif known as the *immunoglobulin fold.* This consists of two antiparallel beta-sheets, one containing five, and the other four, strands, each sheet linked by a disulphide bridge. These are arranged in a Greek-key motif (Figure 4). Four strands from each five-stranded beta-sheet form an eight-stranded beta barrel, and strands from the beta-sheets also form the interface between the V_H and V_L. The two



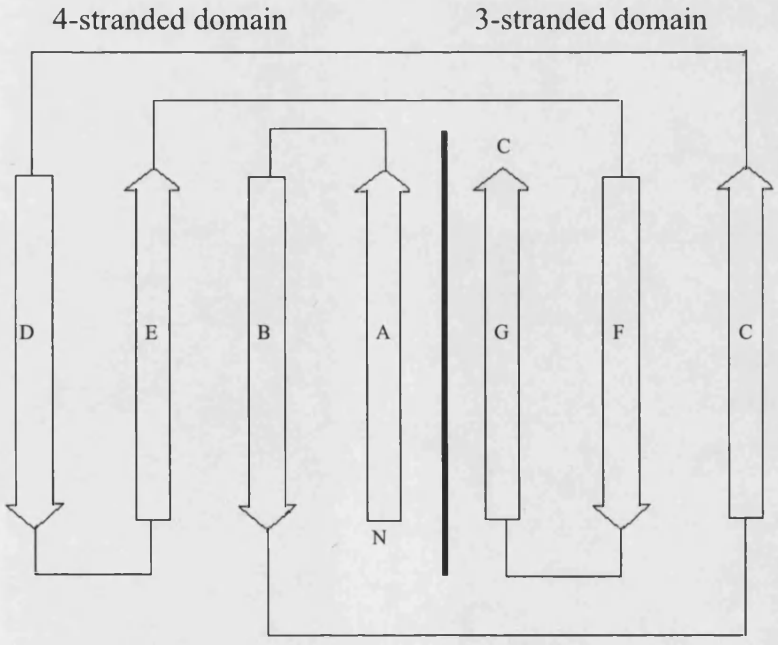


Figure 4.The Greek-key motif of antibody variable (above) and constant (below) domains. The beta-strands are referred to as A to G; the variable regions only contain two additional strands, C and C''. The regions connecting the strands form the loops; those loops which form the CDRs in variable domains are indicated.

domains interact by van der Waals (predominantly), hydrogen bonding and electrostatic interactions.

Protruding from the beta-barrel, and between the beta-sheets in the sequence, are six hypervariable loops, or *complementarity determining regions (CDRs)*. As the terms suggest, these are the most variable parts of the Fv in sequence, and form the surface which actually binds antigen. Their shape also varies depending on sequence and length. The CDRs of the light chain are referred to as L1, L2 and L3, whereas those of the heavy chain are H1, H2 and H3. The remainder of the Fv is much more conserved in sequence, and is known as the *framework* (see Appendix 3).

1.3 Genetic basis of antibody diversity

The first theory on antibody diversity was Ehrlich's (Ehrlich, 1897; translation 1957) in which an immune system cell was thought to contain on its surface antitoxin receptors (equivalent to antibodies- Ehrlich did not know about antibodies as such) for all types of antigen, and secretion of the complementary receptor to an antigen was stimulated by the antigen binding to it. However, this theory became untenable when it became apparent that many new, non-peptide chemicals, could elicit an immune response, as it was considered impossible that there could be enough genes by natural selection for receptors to all these antigens. A later theory, which also

went out of acceptance, was the instructive theory (Breinl and Haurowitz, 1930; Haurowitz, 1952; Mudd, 1932; Alexander, 1931; Pauling, 1940) whereby a flexible antibody molecule is induced by antigen to form a complementary binding site.

The currently accepted theory is selective (Dreyer and Bennett, 1965; Tonegawa, 1983) whereby one B cell produces only one type of antibody when stimulated by the corresponding antigen. There is not one gene for every possible antigen, however; as mentioned from the Ehrlich theory, the rate of mutation would not be sufficient for this result. Instead, diversity is produced by two mechanisms:

a) Different combinations of segments of genes code for sections of the variable region. Different mechanisms operate for λ and κ light chains, and heavy chains, and are discussed below.
b) Different permutations of the rearranged light and heavy chain genes are possible. There are approximately 10,000 of each, leading to around 100 million possible combinations.

λ *chains*: The diversity here is relatively restricted. There are three genes which code for the λ chain: the C gene for the C_L domain, the V gene for residues 1-95 of the V_L domain and the J (joining) gene for the remainder of V_L. There are four V genes, each with its own J gene, and two C genes, which can combine with any V gene.

κ *chains*: Much more diversity is available with κ light chains. In the mouse, there are 250 V genes and 4 J genes, and any V gene can combine with any J gene.

Heavy chains: The V_H domain is more diverse still than the V_L of κ light chains. This diversity is produced by (in the mouse) about 250 possible V genes, 4 possible J genes and in addition, between the two, 10 possible D (diversity) genes. Any combination of the V, D and J genes can be used, leading to around 10,000 combinations. The CDR-H3 is encoded for by the end of the V gene, the entire D gene and the start of the J gene, leading to its particular variability. Further diversity in H3 is produced by terminal deoxyribonucleotidyl transferase which inserts extra nucleotides between the V and D segments.

Recent work has suggested that some of the concepts of the instructive theory may actually operate in certain situations, although this is unlikely to be a widely used mechanism. Holmes and Foote (1997) determined the crystal structure of the Fv fragment of humanised (see later) anti-hen egg white lysozyme antibody, HuLys. This molecule contains CDRs from the mouse anti-lysozyme antibody D1.3, heavy chain framework from the human myeloma protein NEW, and the light chain framework was based on consensus sequences similar to the Bence Jones protein REI. Unexpectedly, the frameworks of HuLys were nearer in conformation to those of D1.3

than either NEW or REI. The effect was strongest at the CDR-framework interface, suggesting that CDRs can induce structural changes in framework residues. This, together with other work by Foote (Foote and Milstein, 1994) suggests that on antigen binding, first the CDR and then the framework could change conformation under certain circumstances; an induced fit mechanism.

1.4 The need for modelling

At present the number of known sequences, both of antibodies and of proteins as a whole, far exceeds the number of known structures that have been solved by X-ray diffraction.

Therefore, a faster techique than X-ray diffraction to solve protein structures is needed - *protein modelling*. This is a technique whereby the structure of a protein can be computed from a sequence using theoretical methods. However, modelling at present does not approach the accuracy of the X-ray method: for a model to be considered accurate it needs to be within 2.0 angstroms (Å, m^{-10}) root-mean-square deviation (RMSD; see Appendix 1) of the X-ray structure. The following section looks at the present state-of-the-art in protein modelling.

1.5 Protein modelling

Overview

Essentially there are three stages of modelling (Sternberg, 1996): sequence analysis to identify homologous known structures to a sequence with unknown structure, secondary structure prediction and tertiary structure prediction. Each stage is summarised and then presented in more detail below.

i) Sequence analysis

The initial stage is to compare the sequence of the protein to be modelled with other sequences in the database, to see if any are homologous by sequence. Having done this, multiple sequence alignment with the homologous sequences takes place. This is useful in identification of functionally or structurally important conserved motifs. Multiple alignments give increased accuracy for secondary structure prediction and subsequent tertiary modelling.

If no homologous sequences can be found, the probable function can be guessed at by identification of previously characterised sequence motifs. This has the disadvantage of assuming that a common function among a range of proteins will have a common

structure. This is not always the case - often the local structure is the same, such as a loop joining a beta-strand and alpha-helix, but the overall fold is different.

ii) Secondary structure prediction

Attempting to predict the secondary structure of a sequence can give more information when predicting the tertiary structure (see below). A number of early algorithms were developed, such as Chou/Fasman (Chou and Fasman, 1974), Garnier-Osguthorpe-Robson (Garnier et al, 1978) and identifying hydrophobic residue pattern identification (Lim, 1974).

Current methods fall into two groups:

a) Computer-based algorithms such as an extension of Garnier-Osguthorpe-Robson (Levin et al, 1993; Garnier et al, 1996), neural networks (Rost and Sander, 1995) and a nearest neighbour approach (Salamov and Solovyev, 1995). An accuracy of 70% in the prediction is typical.

b) "Expert examination" to identify patterns of hydrophobic residues. The sequence is aligned with those of known structures to give an initial idea which residues form the protein core, followed by examination of the core residues to see if they form patterns typical of

certain secondary structure elements, such as an i, i+4 pattern for hydrophobics in an alpha-helix.

iii) Tertiary structure prediction

Essentially there are two approaches, homology and *ab initio* approaches.

The former usually consists of the following stages:

- Find a group of suitable known structures which are homologous to the sequence of the unknown structure and perform sequence alignment.
- Identify the main chain segments expected to be structurally conserved between the known and unknown. Use the most homologous to model the framework of the unknown.
- Loop modelling, using a database of loops from all proteins
- Build the sidechains
- Energy minimisation.

These are discussed further below. If alignments cannot produce homologous structures, the sequence can be 'threaded' through a range of folds to assess its suitability (see below). The most common *ab initio* procedure involves a systematic sampling of conformational space (see below).

Sequence analysis

This has two functions: first, to align the sequence of an unknown with known sequences as part of the modelling process, and second to identify possible functions by comparing with known sequences.

For either function, amino acids in the two sequences need to be compared. Various methods have been used such as genetic code comparison (Fitch, 1966; Cohen et al, 1981) and chemical similarity (McLachlan, 1972; Feng et al, 1985) but the most frequently used method is the *Dayhoff mutation matrix* which scores the match between two residues based on observed mutation frequencies (Dayhoff et al, 1978). The matrix has been updated by Jones et al (1992) to account for the large increase in determined sequences; originally, many substitutions were not observed at all and so suitable weights were determined indirectly.

The Dayhoff matrix can be used to either compare or align the sequences, which is simple in principle, but in practice complications are introduced by the occurrence of insertions and deletions. A technique known as *dynamic programming* (Needleman and Wunsch 1970, Smith and Waterman 1981) is used to deal with these. This has the disadvantage of being rather slow, although some algorithms, such as "Fasta" (Pearson and Lipman, 1988) and "Blast" (Altschul et al, 1990) have been used to increase the speed. Another recent

development is the use of 'evolutionary trees' to find sequences related by evolution (Goldman et al, 1996).

i) Sequence motifs

If no homologous sequences can be found, the probable function can be guessed at by identification of previously characterised sequence motifs. A database, PROSITE (Bairoch, 1991), consists of various structural motifs and the sequences required to form them. A number of programs search the database to allocate motifs to parts of an unknown sequence, such as MacPattern (Fuchs, 1990) and GCG (Devereux et al, 1984). A similar approach is taken by the I-sites database (Bystroff and Baker, 1997).

An important ongoing area of investigation is characterisation of new motifs. The patterns in PROSITE have been determined by inspection; automatic methods are needed to quickly search the large sequence database. A potential problem is that groups of proteins with similar function may have different motifs with only small regions of conserved sequence round the active site.

ii) Methods of characterisation of new motifs

a) Global sequence alignment methods: multiple sequence alignment is used to obtain regions with more than average homology and consensus patterns are then constructed.

b) Statistical analysis of sequences: These methods relate the occurrence of particular sequences with that expected by chance. Sequences which occur more frequently than expected are likely sequences for biological importance.

c) Sequence separated motifs: Often a motif is not made up of continuous sequence, but of widely spaced residues. A group of methods search for motifs of sets of specific residues types separated by specific distances, for example 3 residue types separated by 2 specific distances. If such motifs occur frequently they are likely to be of importance (Posfai et al, 1989,Smith et al, 1990).

d) Deduction from similar conformations: Oliva et al (1997) placed protein loops into classes based on the surrounding secondary structure, the relative geometry of the surrounding secondary structure and the loop torsion angles (torsional clustering was performed). Members of each class were found to have a similar conformation,

providing an automated alternative to visual inspection. Examination of the hydrogen bond and hydrophobic interactions each cluster revealed a number of new motifs for loops.

Secondary structure prediction

As already seen, predicting the secondary structure of a sequence can give important information for use in tertiary modelling. A number of different approaches have been used.

i) Chou/Fasman method

This method (Chou and Fasman, 1974) is based on the statistical propensities for each residue type to form an alpha-helix or beta-strand. These propensities are used to classify each residue into one of 6 classes which each have differing probabilities of forming a helix or strand. The classes are then used to find likely helices and strands in the sequence.

It has the advantage that it is simple, but has the limitation that it only considers the propensities of individual residues rather than whole sequences. Also it is not based on any underlying physical or chemical theory, reference to which can be valuable in discriminating different possible structures.

ii) Garnier-Osguthorpe-Robson method (GOR)

This method (Garnier et al, 1978, 1996) is based on the idea of considering the residue sequence and the secondary structure sequence as two messages related by a translation process, which is examined using information theory. The actual folding is regarded as a 'black box'defined by the observed relation in known structures between the input sequence and output secondary structure. Essentially the information that the input carries about the output is worked out. It has the advantage over Chou-Fasman that it is more theoretically sound, but still does not include any physical or chemical theory - treating the folding process as a 'black box' will not easily lead to an understanding of the physico-chemical principles that guide protein folding.

iii) Other knowledge-based approaches

A number of knowledge-based approaches more sophisticated than the Chou and Fasman method have been developed recently. These include neural networks, which "learn" the connection between sequence and structure (Rost and Sander, 1995; Qian and Sejnowski, 1988); the nearest-neighbour method of Salamov and Solovyev (1995) which predicts the secondary structure of the central residue of

a test segment based on that of homologous segments in proteins of known structure; the method of Zhu and Blundell (1996) which obtains the propensities for each amino acid type to be at each position of a helix or strand and compares the sequence with this profile; and the similar method of Frishman and Argos (1996) which uses the propensities of different residue types to be within certain hydrogen bond patterns (such as those found in helices and strands). The overall success rate of the various methods here varies from 65-80%, depending on how much is already known about the protein to be modelled.

iv) Methods using chemical and physical theory

These methods include those of Lim (Lim, 1974), Cohen (Presnell et al, 1993; Cohen et al, 1986), and King and Sternberg (King and Sternberg, 1990; Muggleton et al, 1992). A set of rules are formulated which relate sequence to the secondary structure they are likely to form by theoretical considerations. For example, an alpha helix is likely to be formed if there are hydrophobic residues at residue i,i+3,i+4,i+7,i+8,.... This method has the disadvantage that database observations are not used, which can be as much a contribution to a good prediction as chemical or physical rules.

Homology modelling

In homology modelling, the tertiary structure is built up, fragment by fragment, from homologous known structures.

i) Fragment assembly

Each framework fragment, e.g. an alpha-helix, beta-strand or beta-turn, is typically modelled from a different structure, either the most sequence-homologous for that region (which is assigned using sequence alignment), or an average structure made up of several homologous structures (e.g COMPOSER: Blundell et al, 1987, Blundell et al, 1988). In COMPOSER, the c-alpha atoms of the homologues are superimposed, the contribution of each being weighted by sequence similarity, and the average position obtained used for the framework.

This leaves the problem of modelling the less well defined loops, which are highly variable even within the same protein class. Typically, a database of loops, either from the same structural class or from the entire PDB, is examined for loops of the same length as the unknown. Additionally, some form of distance constraint is used to limit the loop shape to a suitable one, such as c-alpha to c-alpha constraints (Chapter 2) or end-to-end constraints (such as in COMPOSER; Figure 5).

constraints (Chapter 2) or end-to-end constraints (such as in COMPOSER; Figure 5).

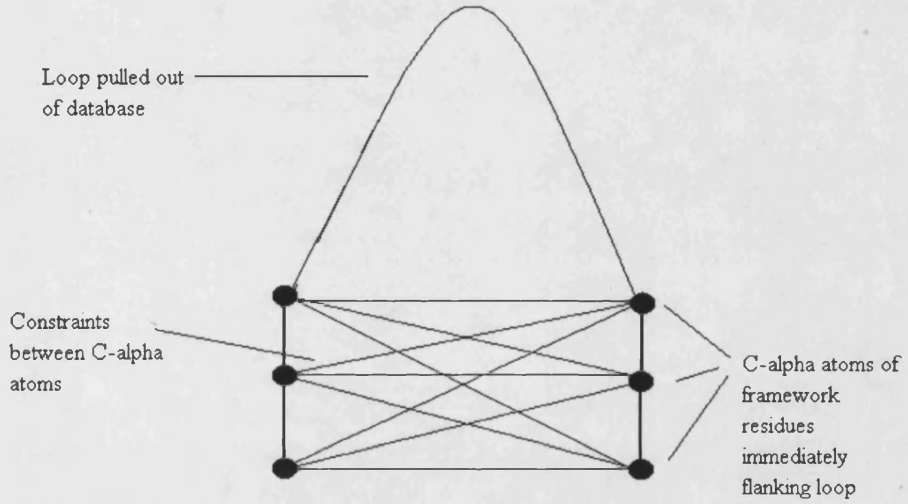


Figure 5. End-to-end distance constraints. A series of constraints are defined between the C-alpha atoms of the residues flanking the loop to be built, which that the framework residues flanking loops pulled out of the database must match.

Ideally, the corresponding loop in a homologous protein should be used, as these are likely to be most similar in conformation. If one is present but it is a different length, the torsion angles can be 'tweaked' to accommodate the insertion or deletion. If no corresponding loops are present, a loop from another class of protein will be needed. The alternative loop construction methods include sampling conformational space (e.g. Bruccoleri and Karplus, 1987), which is discussed at greater length in Chapter 2, or distance restraint methods (see below).

ii) Distance restraint methods

These methods use restraints such as interatomic distances, derived from homologous structures, to construct a model. They can also be used to obtain a model from an NMR structure, using NMR-derived restraints (Podlogar et al, 1997).

Examples include the method of Havel and Snow (1991), and MODELLER (Sali and Blundell, 1993), where a 3D model is derived by optimally satisfying restraints from known structures homologous to the sequence being modelled. The features to be restrained are residue and inter-residue properties, and include solvent accessibility, secondary structure, hydrogen bonding, c-alpha to c-alpha distances, main chain N-O distances and torsion angles. The restraints are

expressed as probability density functions and are obtained from the observed examples of the protein class. Geometrical methods are used to actually calculate the atom positions.

A graph-theory approach is taken by Samudrala and Moult (1998). Each possible conformation of a residue is represented using a 'node' in a 'graph'. Each node is given a weight based on the interaction between its sidechain atoms and the local main chain atoms. 'Edges' are then drawn between pairs of conformations consistent with each other (i.e. no steric clashes) and the optimal sets of 'cliques': completely connected nodes (i.e. conformations) obtained with a 'clique-finding' algorithm. The cliques with the best weights represent the optimal combinations of the various main- and sidechain possibilities. The algorithm can be used in a homology modelling scenario to build sidechains or regions of main chain (such as loops).

iii) Threading

Threading is a technique used to align a sequence with a protein fold, when the tertiary structure is unknown. The sequence is 'threaded' through a number of protein folds, one residue at a time (Figure 6). For each sequence-fold alignment, a score is computed. This is done by assessing the fitness of either single residues, or pairs, for the environment, using such variables as accessibility, secondary structure, and for pairs, distance apart and separation along the chain

(e.g Sippl 1990, Abagyan et al 1994, Willmanns and Eisenberg 1995, Godzik et al 1992, Ouzonis et al 1993, Nishikawa and Matsuo 1993, Jones et al 1992, Gracy et al 1993, Bryant and Lawrence 1993; for review, see Fetrow and Bryant 1993). Once again, dynamic programming is needed to account for insertions and deletions. The fitness scores are calculated using the frequencies of that residue or pair in that environment from a database of known structures.

For each fold, the alignment is taken as the best scoring match, and then the best overall scoring match over all folds is taken as the fold that the sequence adopts.

Recent developments have improved accuracy. One approach, resulting from improvement in the quality of secondary structure prediction, is to first predict the secondary structure of the unknown, and align it with a secondary structure profile of a fold (Rost et al, 1997; Rice and Eisenberg, 1997). Another is to restrict the folds to those homologous (i.e. the same protein type) rather than just analagous (merely sharing a common fold) to the unknown (Russell et al, 1998) - this is likely to become more important as a greater number of structures of each protein type are solved. The advantage of using only homologous folds came about from a finding that merely-analagous structures had little sequence similarity (Russell et al, 1997).

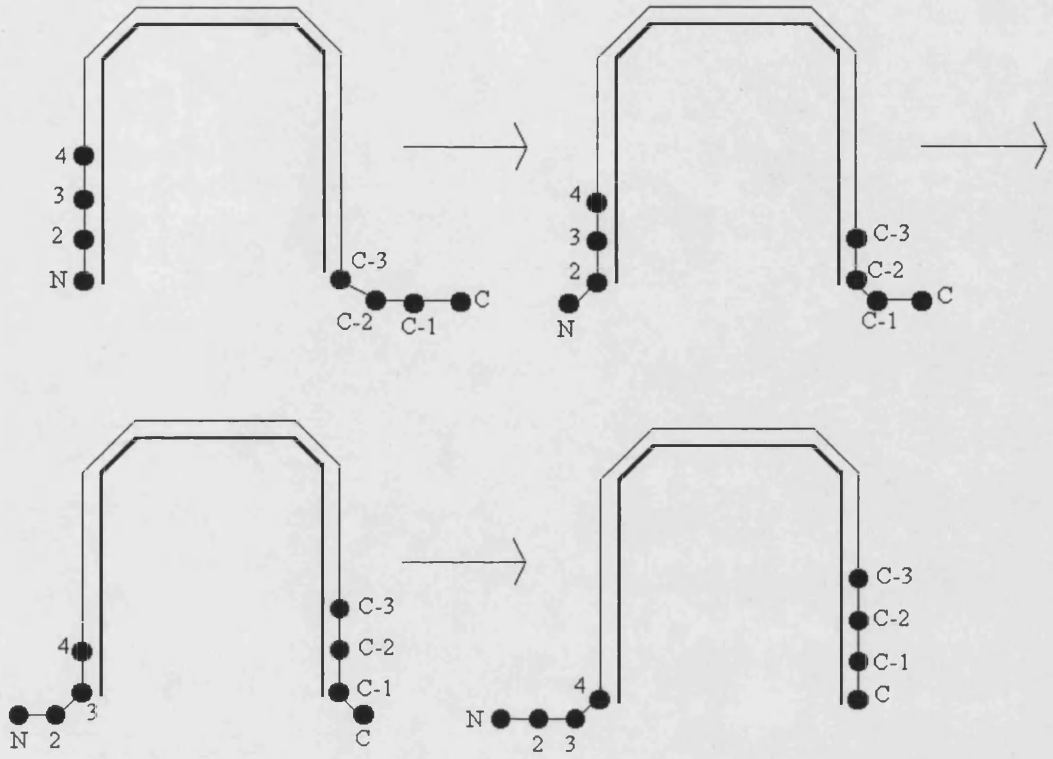


Figure 6. Threading (simplified). The sequence is 'mounted' on the fold with the N-termini of each aligned. The fitness is assessed (see text), and then the sequence is 'threaded' one residue at a time through the fold, assessing the fitness of each sequence-fold alignment, until the C-terminal is reached. The sequence is then 'threaded' in a similar manner through other folds. The overall highest-scoring alignment is noted and the appropriate fold is assigned to the appropriate part of the sequence.

Ab initio methods

Fewer *ab initio* methods exist, as it is clearly more difficult to model a structure without using any prior knowledge. The main approaches include conformational searching, and assembly of secondary structure fragments; an alternative approach is taken by Pedersen and Moult (1997) who use a genetic algorithm to fold the structure (using random torsion angles initially).

Conformational searching

Conformational searching is used in loop modelling to construct a loop residue by residue. The phi/psi torsion angles of each loop residue in turn are sampled, discarding any combination which causes a steric clash with the framework or leads to the remaining residues being unable to span the gap between the growing ends of the loop. An example is the 'CONGEN' algorithm of Bruccoleri and Karplus (1987). This involves a modification of the above procedure: the phi/psi torsion angles of each loop residue, from each end inwards, are sampled, but the middle three residues of the loop are built using a chain closure algorithm which calculates the optimal phi/psi angles to bridge the gap (Go and Scheraga, 1970).

Secondary to tertiary structure

Once the chosen secondary structure prediction method has been applied, the question then remains, how do we combine the secondary structure elements to make a tertiary structure?

There are two main methods:

a) Examine the secondary structures for "packing sites", e.g. groups of hydrophobic residues, and then assemble the elements accordingly in the most energetically 'sensible' way (Richmond and Richards, 1978).

b) A combinatorial approach where structural elements are oriented in all the different possible ways relative to one another. Any combinations which are sterically unacceptable are immediately discounted. (Cohen et al. 1979,1980).

Having then obtained a number of possible structures by either method, further elimination can take place by various tests to make sure that they are sensible structures. For example, tests for whether disulphide bridges (Curtis et al, 1991) and/or metal binding sites (Cohen and Sternberg, 1980) are in the expected place; whether the sidechains are sterically acceptable - they are ignored at the modelling stage (Gregoret and Cohen, 1990); and whether the accessible surface

area of the fold (Teller, 1976) or each residue, based on hydrophobicity (Rose et al, 1985) is typical.

In the more recent method of Jones (1997) simulated annealing is used to find the lowest energy combination of fragments. The potential is a combination of database derived pairwise potentials and solvation energy, and simple terms to favour compact folds but prevent steric clashes.

Energy evaluation

As has been indicated, the loop regions of a model are likely to have a number of possible conformations. A final conformation for each loop must be selected, and the usual means of doing so is to use a potential energy function, such as the Valence Force Field (VFF; Dauber-Osguthorpe et al., 1988), CHARMM (Brooks et al, 1983) or AMBER (Weiner et al, 1984). Prior to evaluation, the model is subjected to an energy minimisation algorithm (Chapter 2), to relieve strain.

Alternatively, knowledge based potentials can be used, in a similar manner to threading, to assess the likelihood of the loop sequence adopting the given structure, by assessing the environments of residues and pairs, and scoring them (e.g Sippl 1990).

i) van der Waals interactions

The van der Waals interactions consist of a repulsive and dispersion (attractive) element. The repulsive forces are due to internuclear repulsions and the Pauli exclusion principle, whereas the dispersion forces arise from small fluctuations of the charge distribution of an atom in the presence of another atom, giving rise to an attraction dipole-dipole interaction, which was shown (Landau and Lifshitz, 1977) to decrease with the inverse 6th power of the interatomic distance.

The repulsive and dispersion elements are usually combined in one expression, the most commonly used being the Lennard-Jones (6-12) potential, which takes the form:

$E=a/r^{12} - b/r^{6}$

r = Interatomic distance

a = Repulsive parameter for this atom pair

b = Dispersion parameter for this atom pair

Different atom pairs have different parameters a and b. They are chosen such that the energy minimum as calculated by the equation agrees with the experimentally observed minimum.

A number of variations on the Lennard-Jones potential have been formulated. The Buckingham potential (Hirschfelder, 1964), essentially replacing the r^{-12} (r is interatomic distance) term with an exponential, $e^{1/r}$ term, is a somewhat more accurate description of the van der Waals forces than the Lennard-Jones, although it has the disadvantage that at very small r, it becomes negative, so it cannot be used at these distances (see Figure 7). Also it is slower to calculate, as it uses an exponential term.

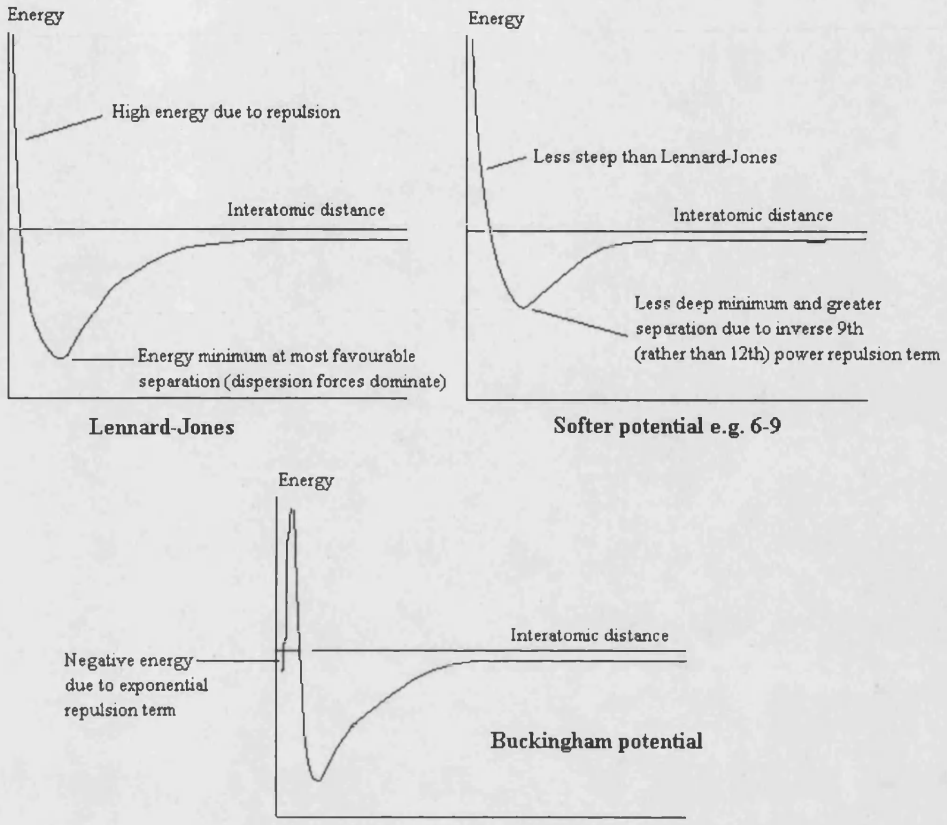


Figure 7. The Lennard-Jones, 6-9 and Buckingham potentials. Note that the "well" is always at the interatomic distance which gives the experimental lowest energy; the parameters are adjusted to force this.

In other investigations (Burkert and Allinger, 1982; Hagler, Huber and Lifson, 1974; Warshel and Lifson 1970) an r^{-9} term has been found to give better agreement with experimental data than r^{-12} for. hydrocarbons, suggesting that a 'softer sphere' model is more accurate. This model lacks the disadvantages of the Buckingham potential.

ii) Electrostatics

Electrostatic energy, the attraction and repulsion between opposite and similar charges, are usually treated by the Coulombic equation:

$$E=q_iq_j/r\varepsilon$$

q_i,q_j = Charges on the atoms

r = Interatomic distance

ε = Dielectric constant

One problem encountered when calculating protein electrostatics is selection of an appropriate dielectric constant. Usually a low value is assigned to the interior of the protein as it is hydrophobic. A value of 2 is a good representation, rather than the vacuum value of 1, as some charged atoms do protrude into the interior (Gilson and Honig, 1988; David Osguthorpe, personal communication).

Another important problem is how to treat the effect of water. As many atoms are at least partially exposed, the dielectric constant of their environment is likely to be considerably above 2, and will approach 80 (the value for water) in the most exposed examples. Approaches to this problem are discussed in Chapter 3.

With both electrostatics and van der Waals, interactions greater than a set distance (typically 10 angstroms) apart are ignored as they contribute little to the overall energy, and ignoring them affords huge time savings. To avoid a discontinuity, a switching function is often used to smooth the transition from included to non-included atoms (Brooks et al, 1983).

iii) Internal energies

The internal energies comprise the "self" energies of bonds, angles and torsions. For bonds and angles, the energy increases the more the value deviates from the equilibrium value. They are usually modelled

by a simple harmonic potential:

Bonds:

$E = k(r-r_o)^2$

k = Bond stretching force constant

r = Bond length

r_o = Reference bond length (standard value)

Angles:

$E = k(\theta-\theta_o)^2$

k = Angle bending force constant

θ = Bond angle

θ_o = Reference bond angle (standard value)

This is appropriate in proteins as most of the motions that occur at ordinary temperatures leave the bonds and angles near their equilibrium values, which appear not to vary by large amounts throughout the molecule. For example, the standard peptide bond length used is an accurate representation.

For torsion angles, it has been found (Scheraga 1968) that the hindered rotation about bonds cannot be modelled with sufficient accuracy by the terms so far considered. If the Lennard-Jones parameters are large enough to ensure a barrier for torsional motion, they are no longer a good representation of intramolecular interactions. So a special term (Scheraga, 1968) is needed to accurately model the phases of torsional energy as the bond is rotated:

$$E=V[1+\cos(n\phi-\phi_o)]$$

V = half the barrier height between torsional minima

n = the periodicity of the torsion (number of minima)

ϕ = Torsion angle

ϕ_o = the reference torsional angle (typically 0 or 180 degrees)

iv) Atom representation

A *united atom* representation (e.g. Dauber-Osguthorpe et al, 1988; Brooks et al, 1983) is a useful means of simplifying the energy calculations. In this representation, only polar hydrogens are included, and the atomic charges on atoms adjusted accordingly. The main advantage of this approach is that calculation times are much reduced, since about half of all protein atoms are hydrogens (H), and the most important interactions involving H atoms are those involving polar H atoms, that is, salt bridges and hydrogen bonds. Other justifications are that the H positions are not usually available in X-ray structures and must be generated from the positions of the other atoms, and also motions involving non-H atoms are separated from H stretching motions, so that removing one type should have only a small effect on the other (Wilson et al, 1955).

The united atom representation has the disadvantage that there are occasionally interactions involving non-polar H atoms, one example being the aliphatic H atoms of a lysine sidechain interacting with the negative charge cloud above a benzene ring of an aromatic residue. Dauber-Osguthorpe et al (1988) obtained somewhat higher RMSD deviations using the united-atom representation compared to all atoms, on modelling the dihydrofolate reductase/trimethoprim complex using energy minimisation (Chapter 2) with the VFF force

field. However, overall the speed advantage of the united atom representation appears to recommend it.

1.6 Antibody modelling

Antibody modelling has an advantage over protein modelling in general. Only the Fv needs to be modelled (the constant region being conserved), and the majority of the Fv itself, the framework, is very conserved in structure between different antibodies, more so than members of many other protein classes. Therefore, the framework can be modelled based on the most sequence-homologous known framework, and in addition, 5 of the 6 CDRs (all except H3) frequently fall into one of between 2 and 6 *canonical* classes, a set for each CDR (Chothia and Lesk, 1987). Members of a canonical class all have approximately the same backbone conformation. This is determined by the loop length and the presence of a number of key residues, both in the CDR and the framework, which hold the CDR in a given conformation by hydrogen bonding, electrostatic and hydrophobic interactions. So, to model an unknown CDR, the sequence is examined, the appropriate canonical class assigned,and the most sequence-homologous known CDR used. For each loop except L2, a few examples fall outside existing canonical classes, and, along with the H3 loop, must be modelled in other ways (see below). However, it may be possible to determine further canonical classes as more crystal structures are solved.

The H3 loop is more difficult to model, as its conformation varies between structures. There are essentially two approaches: knowledge-based methods, such as database searching, where the closest matching database loop (either from antibodies, or from the entire Brookhaven Protein Data Bank (PDB; Bernstein et al, 1977)). in sequence and length is used as the model, or *ab initio* methods, such as the CONGEN conformational search (Bruccoleri and Karplus, 1987).

History of antibody modelling

i) Homology modelling

The first attempt was by Padlan et al. (1976) who made three assumptions:

i) the VL/VH framework of the unknown structure was the same structure as that of the known one,
ii) the quaternary (interdomain) association was the same in both cases,and
iii) the CDRs of the unknown structure were assumed to have the same backbone conformations as other CDRs of known conformation with the same number of amino acids. Sequence similarity was taken

into account when building loops for which there was no starting model, that is, no database loop with the same number of amino acids.

Thus, the framework was modelled on that of the known antibody and the CDRs on known CDRs with the same number of amino acids.This work was extended by de la Paz et al (1986) and subsequently by the canonical class method mentioned above which gives more accurate results since canonical class, and hence conformation, depends on the presence of key residues as well as the length.

ii) Ab initio methods

The first attempt at an *ab initio* method was by Stanford and Wu (1981) who modelled an antibody combining site on the basis of amino acid sequence and steric considerations. They assumed the framework was the same structure as another known antibody, and constructed models of the CDRs by obtaining backbone dihedral angles for tripeptides from known protein structures (usually beta sheet proteins), and imposed them on the segment being reconstructed, according to its sequence. The angles were allowed to vary from the initial value by up to 30 degrees in five degree intervals. The large number of resulting structures was reduced by imposing the conditions that the modelled CDRs should fit onto the assumed

framework structure and that non-bonded atoms should not come within a distance of each other closer than the minimum allowed contact distances based on van der Waals radii. This conformational search approach has been updated by the use of CONGEN (Bruccoleri and Karplus 1987).

The order in which each CDR is modelled is also important to consider. For example, if a structure has canonical and non-canonical loops (the former modelled by the canonical class method, the latter by CONGEN), the non-canonical loops should be modelled after positioning the canonical loops. This is because the reasonably accurate canonical structures should be in place to influence the range of conformations generated for the non-canonical loops through short-range interactions.

There are, however, problems with the CONGEN procedure. For instance the conformations closest to the crystal structure are not necessarily the lowest energy conformations. Even so, in the example discussed above(McPc603), each of the six loops was very closely matched by at least one of the calculated conformations, and in some instances by low energy ones (Bruccoleri et al, 1988).

A different approach to modelling was taken by Fine et al (1986); this involved generating a large number of random conformations for the CDR backbone, which were required to fit onto the framework with the correct geometry. Random phi and psi values were assigned to each angle and these were minimally adjusted by an iterative

procedure to produce the desired fixed-end conditions.After subjecting the structures produced to molecular dynamics and energy minimisation, the lowest-energy structures were taken. Side-chains were then added and energetically-favourable conformations obtained by varying side-chain torsional angles.

Current methods

Current methods of antibody modelling other than those from this laboratory, have generally taken the homology approach, such as in the methods of Pulito et al. (1996), Eigenbrot et al. (1993) and Barry et al. (1994). Pulito modelled non-humanised and humanised variants of an antibody to predict the structure (which was unknown), whereas Eigenbrot and Barry tested the modelling procedure by modelling known structures. Eigenbrot modelled three variants of humanised anti-p185 antibody 4D5, and Barry modelled three anti-DNA antibodies.

In these methods, the most homologous framework from the antibody database is picked, and canonical CDRs modelled on known CDRs of the same canonical class. (Eigenbrot et al. use a slightly different approach for modelling the framework: a number of known structures are 'averaged', followed by energy minimisation to relieve the strain caused by 'average' bond lengths and angles). For the H3 loop, the antibody H3 most closely matching in length and sequence

is used. Deletions are handled by removing the residue and rotating the phi/psi angles of the two surrounding residues to enable a join whilst conserving geometry as far as possible, and insertions essentially the reverse of this. Certain 'canonical-like' key residues can also be taken into account when modelling H3: for example, a salt-bridge forms between residues 219 and 235 (see Appendix 3 for numbering convention) if these are Arg and Asp, respectively (Rees et al, 1996). So if the unknown sequence has Arg and Asp in these positions, the known H3 chosen is one which also has Arg and Asp here (unless the length is very different). Finally, energy minimisation is performed on the structure.

During the modelling process, the framework and CDRs typically come from different crystal structures. A grafting process therefore takes place: when the loops are pulled out of the database, two or three framework residues on each side are also included, and these are fitted onto the corresponding residues on the template framework.

The program ABGEN (Mandal et al., 1996) has automated the homology process described above, with the pre-minimisation stages completed in 6 minutes. These methods (ABGEN and also the work by Eigenbrot and Barry) predict the framework and canonical loops accurately, with global backbone RMSD (see Appendix 1) for these sections less than 1.5Å, but the H3 loop is of much more variable quality (from 1.0 to 4.0Å), due to its greater variability. The alternative approach involves conformational search (Martin,

Cheetham and Rees 1989,1991; Pedersen et.al., 1992), and this is discussed with the AbM program (Chapter 2).

1.7 Why model antibodies?

Antibody modelling (predicting the structure from the sequence) has a number of uses. If an antibody is designed against a certain receptor known to occur on the surface of a known infective agent, then modelling is a method whereby it can be verified whether the designed sequence will lead to the expected structure. Also, for an antibody of known sequence and unknown structure, the mode of antibody/antigen binding can be elucidated.

Another application is in the technique known as *humanisation*. It is easier, and with fewer ethical problems, to produce antibodies against a particular disease in mice than in humans. However, this has the disadvantage that mouse antibodies themselves would be capable of producing an immune response. To find a way round this problem, the concept of humanisation by CDR grafting (Foote and Winter, 1992; Reichman et al, 1988) was put forward. Since the hypervariable loops are involved in antigen recognition, antibodies can be raised in the mouse, the hypervariable loops spliced off and grafted onto a human framework. Therefore, an antibody is produced for which the specific part (hypervariable loops) can be produced easily in the

mouse, and which contains a human Fv framework and constant regions, which are non-immunogenic. Modelling the humanisation designs is important to ensure that the correct framework-CDR interactions are maintained; framework residues may need to be changed to ensure this.

More recently, humanisation by resurfacing has been put forward (Roguska et.al. 1994, 1996; Pedersen et.al. 1994) in which the Fv surface residues only are altered so that the Fv surface resembles a human antibody. This has the advantage over CDR grafting in that changing framework residues, which can increase antigenicity, is not required, and internal framework-CDR interactions are not disturbed.

1.8 Aims and scope of the thesis

The state of antibody modelling is that the only part of the structure which generally cannot be modelled accurately is the CDR-H3 loop, due to its very high variability and absence of canonical classes. Long H3 loops are a particular problem, due to their likely flexibility in solution. The aim of this work was to attempt to improve the accuracy, as measured by backbone global RMSD with respect to the crystal structure, of CDR-H3 modelling. This was achieved by introducing new algorithms within the program AbM, which use a combination of homology and conformational search approaches. A

secondary aim was to attempt to improve sidechain modelling. Sidechain modelling in the CDRs is a particular problem, due to their flexibility: global RMSD for a loop including sidechains is often above 2.0Å, although exposed residues have variable sidechain positions and may not be able to be placed accurately. The success of these improvements was tested against a set of 8 antibodies whose H3 loops varied in length and for which high resolution X-ray structures had been determined.

CHAPTER 2: CDR-H3 MODELLING WITH AbM

2.1 Introduction

The original AbM modelling procedure

AbM (Antibody Modeller, © Oxford Molecular Group plc, 1992) is a package developed by Andrew Martin and other members of the Rees group (University of Bath), which models antibody Fv regions from light and heavy chain sequences. The original AbM modelling procedure is summarised in Figure 8 and is described in Martin, Cheetham and Rees (1989, 1991).

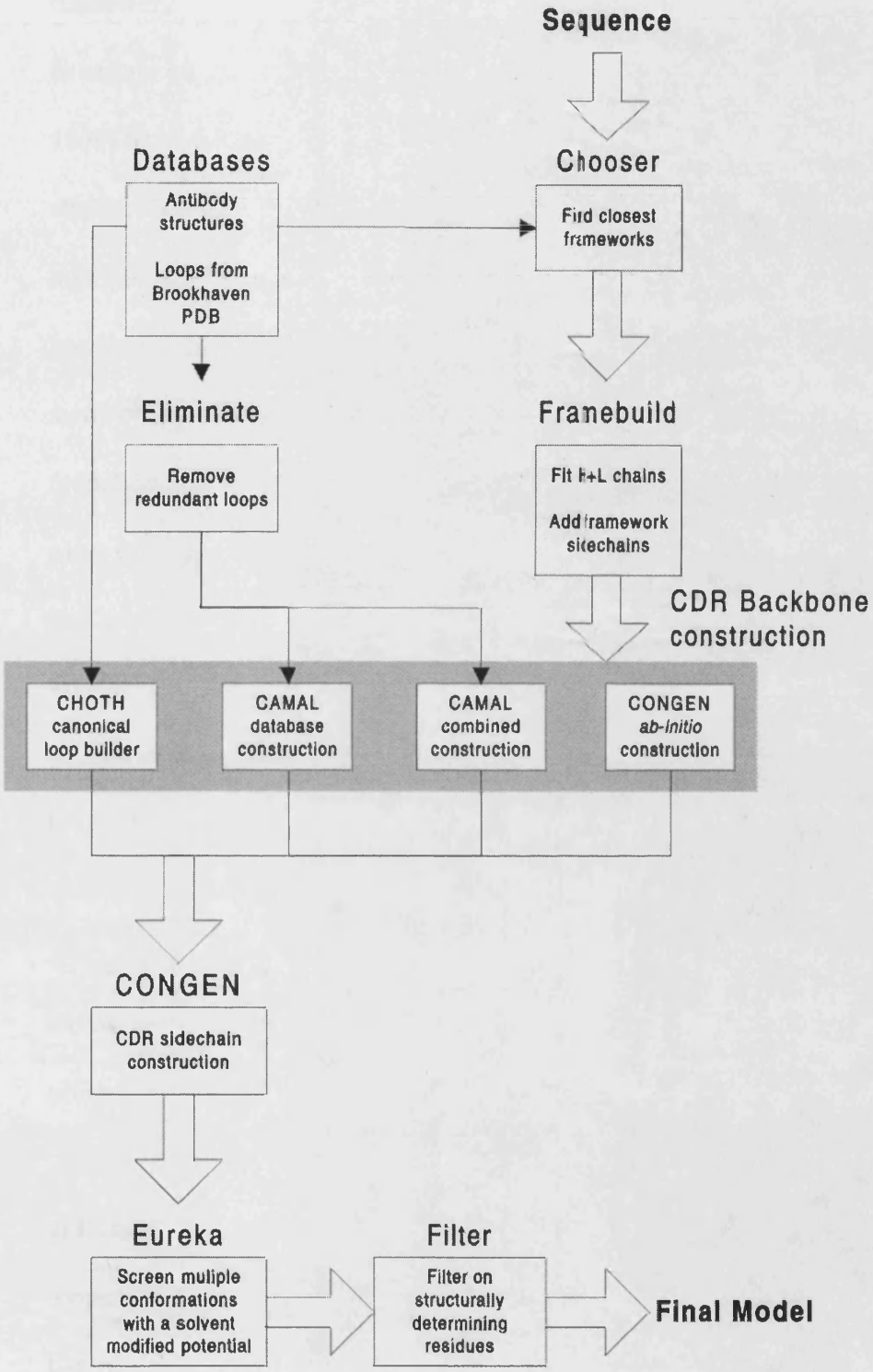


Figure 8. Flow diagram of the AbM antibody modelling algorithm.

The first stage in modelling an unknown Fv is to model the framework. The framework which has the highest sequence homology in certain particularly conformationally-conserved regions (Rees et al, 1996) in an antibody database is chosen for both the light and heavy chains. The sidechains are resequenced by a *maximum-overlap* method, in which the new sidechain is placed with as many atoms as possible in the same position as the old sidechain, while avoiding steric contacts (Snow and Amzel, 1986). The canonical loops are then modelled, choosing the most homologous known structure of the same canonical class, and then resequencing the sidechains by a maximum overlap method. The light and heavy chains are then fitted together.

The next, and most difficult stage, is to model the non-canonical loops. Three possible methods are used to get a range of possible conformations. The nitrogen of the N-terminal loop residue and the C-alpha and carbonyl of the C-terminal residue are not treated as loop atoms, but as framework atoms, and so are excluded from the procedure below.

a) Database search: a search is made in the Protein Data Bank for loops which match the length of the unknown loop and have a set of C-alpha to C-alpha distances (the N terminal C-alpha to the other C-alphas, and the C terminal C-alpha to the other C-alphas; Figure 9) in the range defined by the mean value, plus or minus the standard

deviation multiplied by 3.5. (Three standard deviations includes all values in a normal distribution, but 3.5 is used here due to the small sample size, to account for the possibility of values not yet observed).

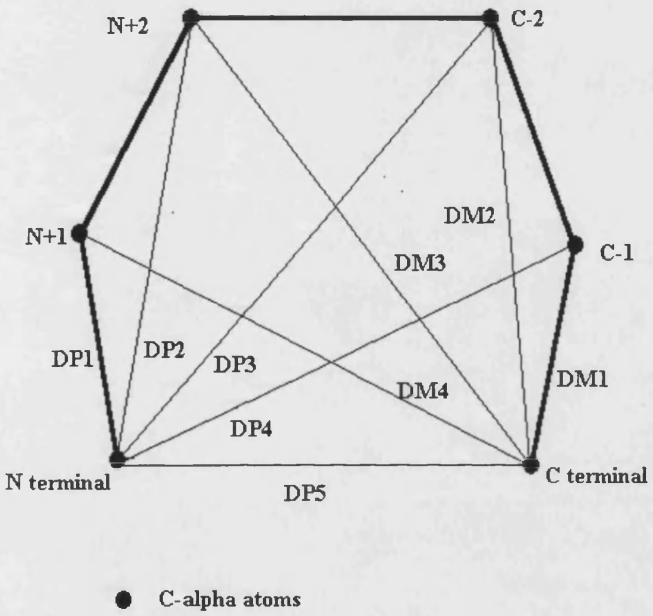


Figure 9. The C-alpha to C-alpha distance constraints search for CDRs. The PDB is searched for correct length loops which have a set of inter C-alpha distances from the N-terminal C-alpha to the other C-alphas (DP1, DP2, DP3, DP4 and DP5 above) and from the C-terminal C-alpha to the other C-alphas (DM1, DM2, DM3 and DM4 above).

b) CONGEN conformational search (Bruccoleri and Karplus 1987):

The phi/psi torsion angles of each loop residue in turn, from each terminal inwards, are sampled, discarding any combination which causes a steric clash with the framework or leads to the remaining residues being unable to span the gap between the growing ends of the

loop. The final three residues of the loop (the middle three) are built by a chain closure algorithm which calculates the optimal phi/psi combination (Go and Scheraga, 1970).

c) The combined algorithm (CAMAL). First a constraint-based database search is performed on all structures in the PDB, then the middle five residues of each loop are deleted. They are rebuilt with CONGEN, the outer two built by conformational sampling, and the middle three built using the chain-closure algorithm. This method (Figure 10) combines the advantages of the speed of the database search with the greater sampling of conformational space of CONGEN. It requires a loop length of at least 7 residues; shorter loops are modelled using either the database search alone, or CONGEN alone.

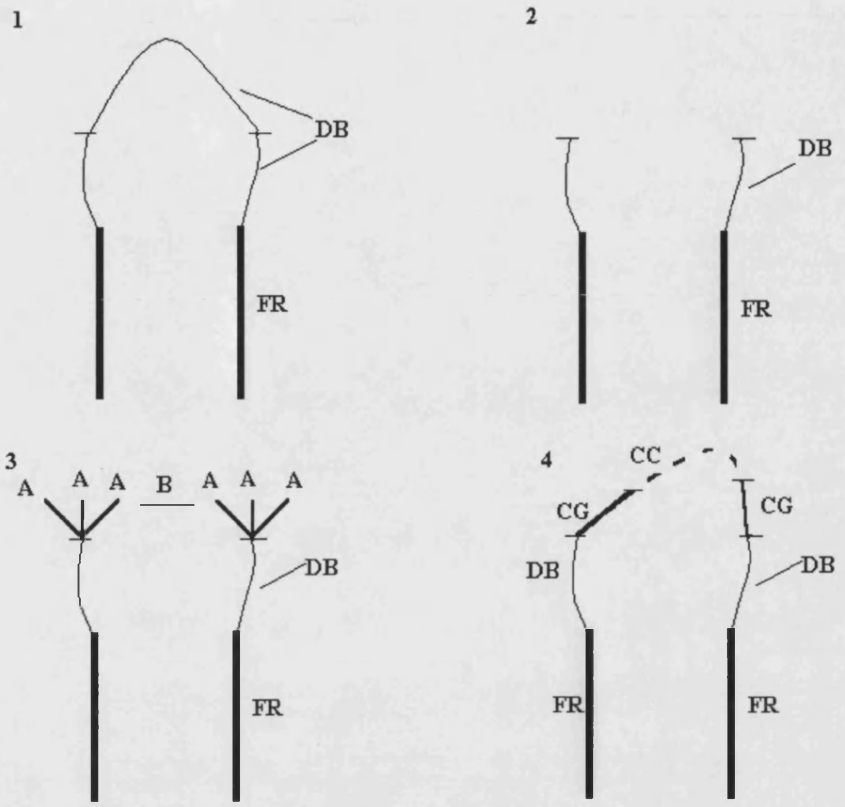


Figure 10. The CAMAL loop-building procedure.
Abbreviations: FR=framework; DB=region built using database search; CG=region built using CONGEN; CC=region built using the chain-closure algorithm.
1 - A loop from the database search is initially grafted.
2 - The middle 5 residues are deleted.
3 - The outer two residues of the middle 5 are rebuilt using CONGEN. The conformational space is sampled, and for each phi/psi combination (A), the distance to the far end of the loop is checked (B) to make sure that remaining residues can span the gap.
4 - The middle three residues are built using the Go and Scheraga chain closure algorithm

Whichever method is used, sidechains are added using the CONGEN iterative method (Appendix 2). The conformations thus produced are energy screened using the implementation of the VFF (Valence Force Field; Dauber-Osguthorpe et al., 1988) within AbM, known as Eureka. This force field consists of terms for bond stretching, angle bending, torsional energy, and repulsive van der Waals (see Chapter 1).

Note that in Eureka within AbM the dispersion term of the Lennard-Jones potential is turned off, so the energy is simply a/r^{12}. In addition, the VFF force field (but not in Eureka within AbM) contains an electrostatic term, as below. The reason for ignoring the dispersion and electrostatic terms in AbM will be discussed shortly.

Finally, the bottom five energy loops are screened using a Structurally Determining Residue (SDR) filter, which scores the conformations depending on whether the torsion angles are typical for that residue type in that residue position, based on the observed patterns in the loops extracted from the database.

Problems with the 'old' *AbM modelling procedure*

Using AbM as a modelling algorithm, the framework and canonical CDRs are modelled accurately, as for homology-only approaches, but non-canonical structures, including all H3 loops, are not always modelled accurately (global RMSD over backbone CDR-

H3 atoms is in the range 1.0 to 4.5Å). Sidechains in particular are less well modelled: in some CDR-H3 residues, the RMSD with respect to the crystal structure sidechain can be as high as 7.0Å.

There are a number of problems which were found to contribute to poor CDR-H3 models. First, Eureka as implemented carried a number of 'bugs'. Most seriously, the bond stretching potential was divided by 2, altering the energy order of the conformations when combined with other terms. Also, the definition of bonding in a histidine ring was incorrect - the delta-hydrogen atom, which should have been bonded to the histidine ring delta-N atom, was bonded to the delta-C atom instead. In actual fact this did not alter the order of conformations because the resulting bond and angle energies are much higher relative to the other terms, but it was an example of how small errors in written code can lead to potential problems in output accuracy.

Second, when database loops were grafted onto the framework, bond angles at the join were allowed to take on high energy values without penalty in order to make the database loops fit. The energy will depend on the spatial match between the framework and the loop, and consequently, low RMSD loops can end up with higher energy than high RMSD loops.

A third problem related to the selection of an appropriate force field. The full VFF forcefield (within Eureka) ignores solvent effects. Previous investigations have shown that this leads to 'collapsed' loops

(Martin et al, 1989). Therefore, a solvent-modified potential was used within AbM, which took the solvent into account in a simple way, by turning off the electrostatic and dispersion van der Waals terms from the VFF potential. This causes the repulsive term to dominate, leading to selection of more 'outward-pointing' loops. In actual fact, it over-compensated, leading to 'ballooning-outwards conformations (Figure 11).

Finally, it was hard to predict the conformations of long loops (greater than 10 residues) in any instance. They are, of course, more likely to be flexible in solution as length increases.

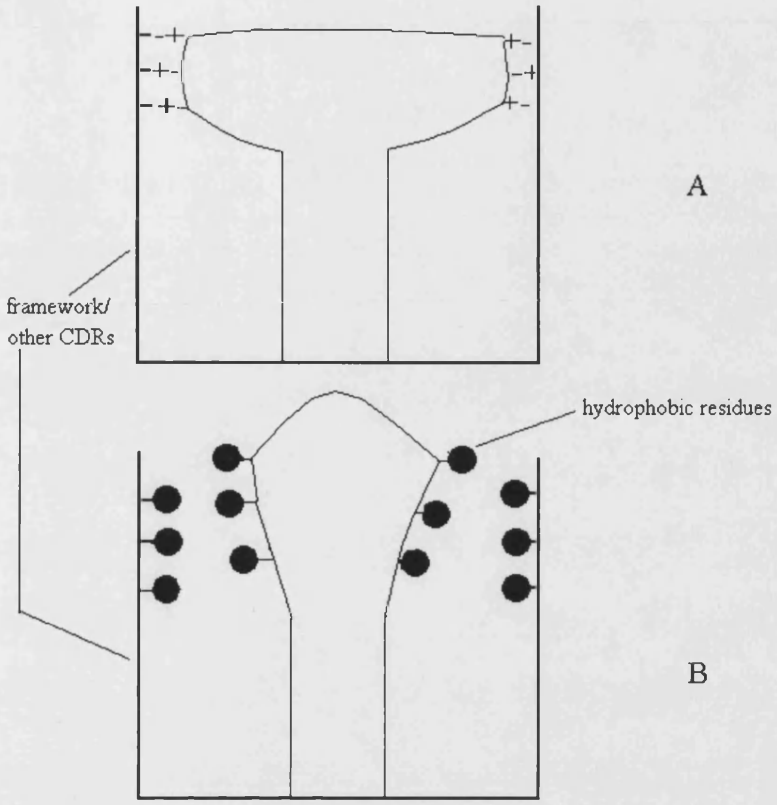


Figure 11. The effect of including and excluding electrostatics in the VFF on the stability of conformations. The "collapsed loop" (A) is stabilised by electrostatic interactions which dominate in the full VFF; in the partial VFF, with no electrostatics, the repulsive van der Waals term dominates,leading to "ballooned out" conformations (B).

2.2 Methods used in assessing the limitations of 'old' AbM

Obtaining the structures

The Brookhaven Protein Data Bank was searched for uncomplexed structures, one for each CDR-H3 length from 5 to 12 residues, which met the following requirements:

a) resolution < 3.0 Å (and preferably R factor < 0.2);

b) no missing residues in any of the CDRs.

If a suitable structure could not be found for a given length, criterion a) was relaxed to include structures with resolution and R factor just outside the stipulated limits. Under no circumstance was criterion b) relaxed.

The following structures were obtained:

Length	PDB code	Antibody identity	Antigen	Resolution (Å)	R value	Reference
5	1bbj	B72.3	Mucin-like glycoprotein from tumour cells	3.1	0.176	Brady et al, 1992
7	1cgs	NC6.8	Nonpeptide sweetener	2.6	0.218	Guddat et al, 1994
8	1mam	YsT9.1	Brucella A cell wall polysaccharide	2.5	0.215	Evans et al, 1994
9	2fbj	J539	Galactan	1.95	0.194	Suh et al, 1986
10	1for	Fab-17	Human rhinovirus surface protein	2.75	0.174	Liu et al, 1994
	1igf	B13I2	C-helix peptide from myohaemerythrin	2.8	0.180	Stanfield et al, 1990
11	1hil	Fab 17/9	Synthetic peptide from influenza haemagglutinin	2.0	0.190	Schulze-Gahnen et al, 1993
12	1igm	IgM Pot	Antibody is IgM in Waldenstrom's macroglobulinaemia	2.3	0.201	Fan et al, 1992

(No 6 residue H3 satisfied the criteria set; to compensate, two 10 residue CDR-H3 structures were selected).

Modelling different length CDR-H3 loops

The CDR-H3 of each structure was modelled with the standard AbM modelling procedure, using the crystal structure for the framework and other CDRs. CAMAL was used for all structures except 1bbj, which has a 5 residue H3 loop, and for which the database search was used. VFF was used as the energy screen, using both the full and solvent-modified potential (except for structure 1hil, for which the full potential only was used); when using the full potential, a dielectric constant of 2 was used for the electrostatic

energy as this was considered more representative of the interior of a protein than the value of 1 used for small molecules (Gilson and Honig, 1988; David Osguthorpe, personal communication). The final structure was selected by energy, and also by the SDR filter for the solvent-modified potential.

Altering the rebuild range

By default, CONGEN in CAMAL rebuilds the middle 5 residues. This is because it had been assumed that this is the most variable part of the H3 loop, and therefore needed the greatest variation in the models in order to maximise the generation of crystal structure-like conformations. Examination of crystal structures with Insight II, however, appeared to show that the most exposed area (and therefore most likely having antigen binding, and hence variable, residues) was towards the N terminus. To confirm this, the relative accessibility of each residue position in all free antibody crystal structure H3 loops was measured.

Relative accessibility = 100(ABS / MAX)

where ABS, the absolute accessibility (the exposed surface area in $Å^2$) is measured using the DSSP method (Kabsch and Sander, 1983)

and MAX, the maximum absolute accessibility is that for the residue in a Gly-X-Gly tripeptide in the helical conformation (Table 1).

Table 1. Relative accessibilities of residues of H3 loops. Residues in bold are exposed (>=30% relative accessibility) and residues in lower case are buried (<30% relative accessibility). N=N-terminal; C=C-terminal.

	N											C
1bbj	s								y	y	g	**H**
1ggb	**E**								g	y	i	**Y**
1hkl	**Y**								**Y**	g	i	**Y**
1cgs	g	**Y**	s						s	m	d	**Y**
1mlb	g	**D**	g						n	y	**G**	**Y**
1mrc	l	**R**	g						**Y**	f	d	**Y**
1mam	d	p	**Y**	g					p	a	a	**Y**
1vfa	e	**R**	**D**	**Y**					r	l	d	**Y**
1plg	g	**G**	**K**	F					a	m	d	y
1kem	w	g	**S**	**Y**					a	m	d	**Y**
2fbj	l	**H**	**Y**	**Y**	g				y	n	a	**Y**
1mfb	g	g	**H**	g	y				y	g	d	y
7fab	n	l	**I**	**A**	g				g	i	d	v
1dba	g	d	**Y**	**V**	n	**W**			y	f	d	v
1for	s	g	**N**	**Y**	p	**Y**			a	m	d	**Y**
1igf	y	s	**S**	d	**P**	**F**			y	f	d	**Y**
1igi	s	s	**G**	**N**	k	**W**			a	m	d	y
1nbv	d	q	**T**	**G**	t	a			w	f	a	**Y**
1rmf	g	g	**W**	**L**	**L**	l			s	f	d	**Y**
1ucb	g	l	**D**	**D**	g	a			w	f	a	y
1fvc	w	g	**G**	**D**	**G**	f	**Y**		a	m	d	**Y**
1hil	r	e	**R**	**Y**	d	e	n		g	f	a	**Y**
1mcp	n	**Y**	**Y**	**G**	**S**	t	w		y	f	d	v
1ngq	y	d	**Y**	**Y**	**G**	s	s		y	f	d	**Y**
1igm	h	r	**V**	**S**	**Y**	v	l	t	g	f	d	s
1vge	d	p	**Y**	**G**	**G**	g	**K**	s	e	f	d	**Y**
6fab	s	**E**	**Y**	**Y**	**G**	g	s	**Y**	k	f	d	**Y**

In view of this observation, the modelling of five of the structures was repeated, with the rebuild regions shifted towards the N-terminus of H3 to reflect the most variable region of the loop, as below. The CONGEN rebuild range is enclosed in square brackets and the chain closure range in standard brackets.

Structure	New rebuild	Original rebuild
1mam	D[P(YGP)A]AY	DP[Y(GPA)A]Y
2fbj	L[H(YYG)Y]NAY	LH[Y(YGY)N]AY
1for	S[G(NYP)Y]AMDY	SGN[Y(PYA)M]DY
1igf	Y[S(SDP)F]YFDY	YSS[D(PFY)F]DY
1igm	H[R(VSY)V]LTGFDS	HRVS[Y(VLT)G]FDS

Energy minimisation

As noted earlier, the conformations generated are poorly grafted to the framework, with high energy bond angles at the join. In addition, conformational search-built loops are also likely to be strained. To combat these problems, energy minimisation can be performed.

Energy minimisation is a technique for lowering the energy of a molecule by small adjustments of the atoms. If an energy 'surface' is imagined where each point represents a different conformation, minimisation aims to locate the minima by finding the gradient at a given point, which is given by the derivative of the energy. There are two methods commonly used, *steepest descent* and *conjugate gradients.*

The *steepest descent* method adjusts the coordinates to move directly downhill on the energy surface, in the direction of the gradient, by a given step size (typically 0.02 of the unit vector in that direction). If the energy is lowered, the step size is then increased. If, however, the energy increases, a minimum has been missed and the step size is reduced, in order to try and find the minimum between the previous two points. The steepest descent algorithm has the advantage of being relatively fast, and is effective where high energy clashes need to be removed, as it effectively finds an area of general low energy. However, it has the problem that it can be slow in finding the absolute minimum due to continual 'jumps' from one side of a minimum to the other.

The conjugate gradients method attempts to solve this problem by considering the previous direction moved when choosing which direction to move next on the energy surface. The first step is straight down the gradient by a given step, as for steepest descent, but subsequent steps combine the gradient with the previous direction moved:

$$D_c = -G_c + (G_c^2/G_p^2)\, D_p$$

where D_c is the direction to move, G_c is the current gradient, G_p is the previous gradient and D_p the previous direction. It can therefore be seen that the previous direction is taken into account to a greater

extent if the gradient has increased, and vice-versa. This is therefore an effective way to locate the absolute minimum, though it is slower than the steepest descent method as more energy calculations are needed.

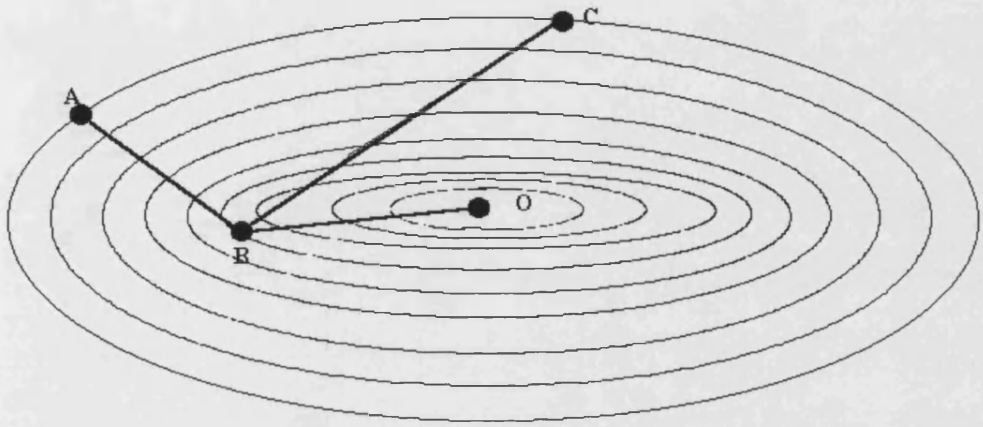


Figure 12. Steepest-descent and conjugate gradients minimisation. The above is an energy surface contour map, with the minimum at point O, and each coordinate on the map representing a conformation. If a conformation starts at point A on the map and is to be minimised, the first step (whichever method) is downhill in the direction of steepest descent from A. If it is assumed that this first step is successful in locating the minimum along this direction (B), i.e. the optimal step size is used, the success in locating the overall minimum differs depending on whether steepest-descent or conjugate gradients is used. The former will now take the direction of steepest descent from B, and the resulting conformation will be somewhere on line BC (depending on the step size). Conjugate gradients, however, will successfully locate the minimum (O) by travelling in a direction combining the current steepest descent and previous search direction.

The method of choice here was steepest descent, since we were more interested in finding a general low energy area on the energy surface, by relieving the areas of high energy, than finding the absolute minimum. In order to determine the optimum rounds of minimisation, a number of conformations from four of the structures, both low and high in energy, were minimised using 5,10,15,20,25,30,35,40,45 and 50 steps of steepest-descent minimisation, with the full VFF potential. Only the H3 loop was minimised since we wished to retain the crystal structure framework. The optimum number of rounds was determined by ranking the structures after each minimisation. The optimum number was the first for which the ranking remained the same after 15 steps (e.g. the ranking after 30 rounds was the same as that with 25,20 and 15 rounds, for all the structures) and the RMSD deviation with respect to the non-minimised structure was less than 1.0Å. This optimum number was found to be 45 rounds.

Therefore, the modelling procedure was repeated for each structure, using 45 rounds of steepest-descent minimisation within VFF. The charges on the H3 sidechains were turned off during minimisation, to prevent the formation of false salt bridges. Following the minimisation, the H3 sidechains were recharged and the conformational energies were recalculated.

Using individual energy terms to screen the loops

One possible source of the poor RMSD/energy correlations observed with AbM (old) is that the overall force field could have been masking some important counteracting effects. For example, some low RMSD conformations may have had favourable van der Waals energies, but some high RMSD conformations, despite having higher van der Waals energies, may have had very low electrostatic energies due to fortuitous positioning of oppositely charged sidechains. In view of this, the individual terms of the VFF were used in turn as a screen.

Each individual VFF component (internals, van der Waals and electrostatics) was used as a screen for the set of energy minimised loops, and the spread of the bottom 200 noted. In addition, in order to see which component correlated best with RMSD, the energy ranking of the components were determined for the 10 lowest overall energy minimised loops for each structure. In structures where all the 10 lowest energy conformations were either of low or high RMSD, conformations slightly further down the ranking were also taken, in order to give a range of high and low RMSD conformations on which to perform the test.

It should be noted here that an additional structure was used from this point on, 1vfa (Bhat et al, 1994). This was a new structure (D1.3;

anti-lysozyme) which appeared in the PDB since the start of the work and had good resolution and R factor (1.8Å and 0.218 respectively).

2.3 Results

Modelling different length CDR-H3 loops by 'old' AbM

Table 2 shows the bottom 5 energy loops for each structure, along with the conformations picked by the SDR filter. The results are rather mixed. Half the structures have both a lowest energy loop less than or equal to 2.0Å in RMSD and a spread of conformations in the bottom 5 generally below 2.0Å RMSD (2fbj, 1for, 1igf and 1bbj) whereas for the other half (1cgs, 1mam, 1hil and 1igm) the lowest energy loop is above 2.5Å in RMSD (above 4.0Å in 1mam and 1hil), and all the bottom 5 conformations are well above 2.0Å in RMSD. These results apply to both the solvent modified and full VFF; the results are almost identical whichever forcefield is used.

In most cases, the RMSD values of the bottom 5 loops are similar, so that nothing can be said about the effectiveness of the SDR filter. In 1cgs the filter picks a conformation of RMSD above 4.0Å over conformations with RMSD below 3.0Å , so is not effective here.

The best modelled H3 is that of the 5-residue H3 loop of 1bbj, where all the bottom 5 conformations have RMSD below 1.5Å. By

Table 2. Comparison of the RMSD of the 5 lowest energy conformations from the modified and full VFF and the modified VFF with the altered rebuild range.
Energies are in kcal/mole; RMSD is in angstroms.* indicates the conformation selected by the SDR filter.
'Conf' indicates the arbitrary CONGEN conformation number; this applies to **all** tables.

1bbj

Mod.			Full			Altered	range	
Conf	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*
2341	2965	1.2	1198	2831	1.3	-	-	-
2151	2967	0.7	2151	2832	0.7	-	-	-
1198	2972	1.3	2341	2840	1.2	-	-	-
2544	2985	1.3	2836	2847	1.0	-	-	-
836	2986	1.0	87	2864	1.0	-	-	-

1cgs

7827	2739	2.8	7828	2563	2.8	-	-	-
7828	2744	2.8	7832	2568	2.6	-	-	-
7832	2744	2.6	7827	2570	2.8	-	-	-
5394	2749	4.3*	8690	2575	4.0	-	-	-
7830	2759	2.8	3663	2581	3.5	-	-	-

1mam

3317	3442	4.6*	3317	3218	4.6	449	3428	4.6
3319	3456	4.6	3319	3231	4.6	447	3432	4.7
88	3462	4.4	3316	3240	4.6	448	3433	4.5
3316	3501	4.6	88	3247	4.4	453	3442	4.6
3321	3513	4.6	3313	3272	4.5	456	3445	4.9

2fbj

64	3266	1.7	64	2979	1.7	3780	3198	2.6
49	3269	1.8	49	2990	1.8	227	3199	1.7
56	3283	1.8*	56	2997	1.8	19340	3199	2.2
47	3296	1.8	47	3007	1.8	3776	3200	2.6
45	3309	1.9	60	3018	1.7	311	3200	2.3

1igf

2037	3361	1.8	2037	2942	1.8	402	3157	3.0
2028	3382	1.9	2028	2969	1.9	444	3168	2.4
2039	3465	1.8*	2039	3041	1.8	438	3174	2.8
2005	3499	2.2	2005	3044	2.2	429	3176	2.7
2010	3556	2.0	2010	3111	2.0	422	3181	3.2

1for

Mod.			Full			Altered	range	
Conf	Energy	RMSD	Conf	Energy	RMSD	Conf	Energy	RMSD
5877	2487	1.5	5880	2182	1.5	3370	2413	2.5
5880	2487	1.5	5877	2182	1.5	4209	2415	1.6
5874	2499	1.6	5874	2195	1.6	4205	2416	1.6
5882	2510	1.5*	5882	2201	1.5	3329	2417	2.7
5864	2528	1.6	5864	2223	1.6	4268	2417	1.6
1igm								
15802	3219	2.6	15802	2818	2.6			
15803	3230	2.6*	15803	2827	2.6	-	-	-
15700	3238	2.7	15764	2828	2.6	-	-	-
15764	3239	2.6	15700	2835	2.7	-	-	-
15664	3265	2.8	15796	2843	2.5	-	-	-

contrast, two of the poorest-modelled structures are 1hil and 1igm, which both have long loops (11 and 12 residues respectively). These long loops are more flexible, changing conformation in solution more than the shorter loops, so there is a greater likelihood of low-energy structures differing considerably from the crystal structure, whereas the conformation of the short loop of 1bbj is largely determined by intra-Fv forces.

Altering the rebuild range

The results (Table 2) show that there is no consistent improvement. In 1mam, high RMSD conformations (above 4.0Å) are again picked in the bottom 5, and in the other three structures a deterioration is seen. In 1for, two conformations above 2.0Å are picked in the bottom 5, compared to none with the original rebuild range; in 2fbj, three are above 2.0Å, compared to none; and in 1igf, all are above 2.0Å compared to two in the original rebuild range.

Energy minimisation

Table 3 shows the bottom 5 energy conformations for each structure. Although there is no clear improvement in RMSD distribution, if the analysis is extended to the bottom 200

Table 3. Comparison of the RMSD of the bottom 5 energy conformations with no minimisation and with 45 rounds of steepest-descent minimisation.
The full VFF is used in both cases; energies are kcal/mole and RMSD is in angstroms.

	No minimisation		Minimisation		
	Conf	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*
1bbj						
	1198	2831	1.3	2544	2661	1.1
	2151	2832	0.7	931	2661	1.5
	2341	2840	1.2	1198	2662	1.2
	2836	2847	1.0	2151	2662	0.6
	87	2864	1.0	389	2662	0.4
1cgs						
	7828	2563	2.8	4129	2414	2.4
	7832	2568	2.6	4375	2415	3.5
	7827	2570	2.8	2952	2419	4.0
	8690	2575	4.0	4402	2419	3.6
	3663	2581	3.5	4383	2420	3.7
1mam						
	3317	3218	4.6	3317	3099	4.4
	3319	3231	4.6	3325	3102	4.3
	3316	3240	4.6	3316	3105	4.4
	88	3247	4.4	3315	3105	4.4
	3313	3272	4.5	3319	3106	4.5
2fbj						
	64	2979	1.7	6173	2835	1.9
	49	2990	1.8	4424	2835	1.9
	56	2997	1.8	6167	2836	1.9
	47	3007	1.8	821	2837	2.4
	60	3018	1.7	6171	2837	1.8
1igf						
	2037	2942	1.8	2028	2692	2.0
	2028	2969	1.9	2039	2695	1.9
	2039	3041	1.8	2037	2698	1.8
	2005	3044	2.2	1989	2700	2.6
	2010	3111	2.0	2001	2701	2.5

1for

No	minimisation		Minimisation		
Conf	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*
5880	2182	1.5	5864	2082	1.5
5877	2182	1.5	5869	2083	1.5
5874	2195	1.6	5880	2083	1.4
5882	2201	1.5	5877	2084	1.4
5864	2223	1.6	677	2085	1.9
1hil					
3164	2331	4.2	3362	2093	4.1
3590	2332	4.2	3557	2093	4.3
3591	2333	4.3	3590	2094	4.1
3362	2333	4.2	3395	2094	4.3
3557	2333	4.5	3164	2097	4.0
1igm					
15802	2818	2.6	15802	2602	2.8
15803	2827	2.6	15814	2603	2.7
15764	2828	2.6	15764	2603	2.8
15700	2835	2.7	15803	2604	2.8
15796	2843	2.5	15809	2604	2.7

conformations (Table 5, Figure 13), an improvement is seen in 1bbj, 1cgs, 1for and 2fbj, a definite improvement in 1mam, no change in 1igf, while 1igm is the only structure to have a slightly worse RMSD distribution. In addition, although an improvement is seen in the spread for 1cgs, it is still poor with few loops below 2.0Å RMSD.

Splitting the VFF into its component parts

The results (Table 4) show that of the lower overall energy loops, whereas in 3 out of 7 cases (1vfa, 1for and 1igf) there is a correlation between RMSD and internals energy, there is not in general a correlation between low RMSD and good ranking in van der Waals energy. There is also is a tendency for the high RMSD loops to have the best electrostatic rankings (e.g. 1cgs, 1for).

Using the individual components of VFF as an energy screen for all loops

The results (Table 5) show that using each individual component rather than the full VFF as the energy screen, does not improve the RMSD distribution of the 200 lowest energy loops. Both internals and van der Waals perform similarly to the full VFF in selection of low RMSD conformations (e.g. 1bbj, 1cgs, 2fbj, 1igf), whereas

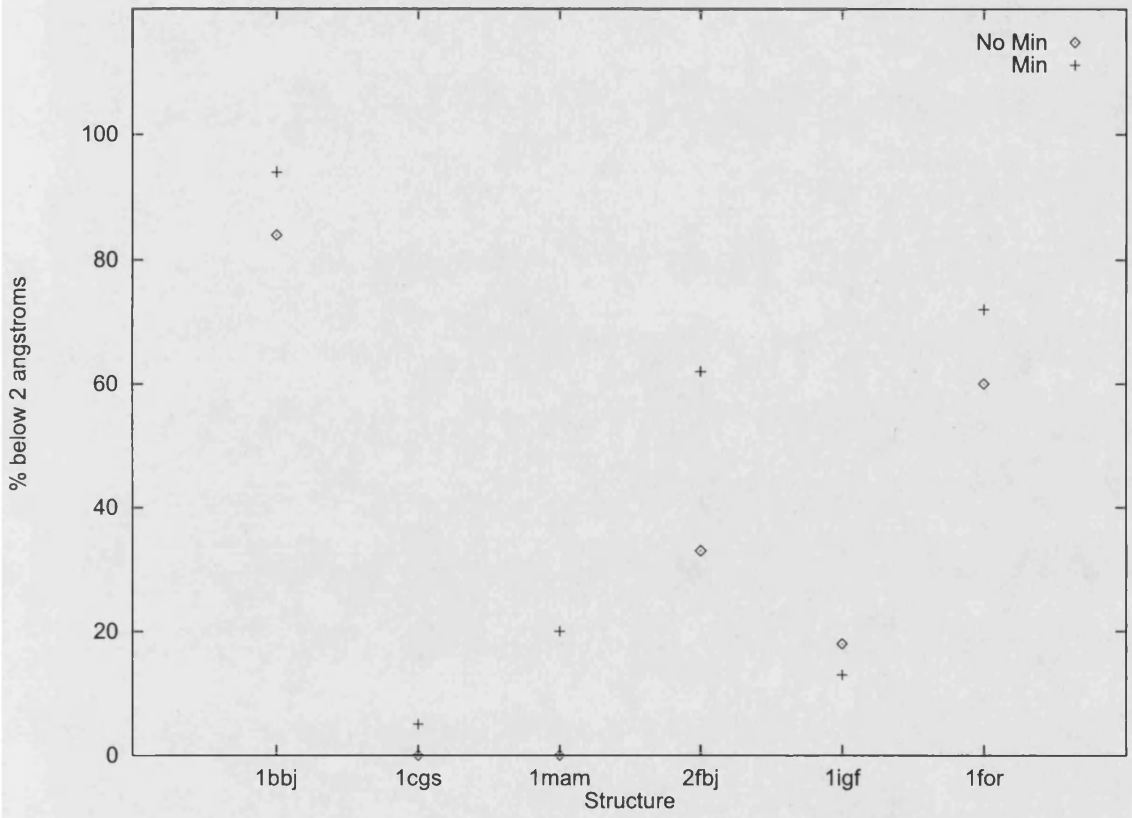


Figure 13. The percentage of the bottom 200 energy H3 conformations below 2 angstroms, VFF non-minimised vs. Minimised.

electrostatics perform decidedly worse, selecting higher RMSD loops in most cases (e.g. 1bbj, 1cgs, 1vfa, 1for).

Table 4. The bottom 10 minimised conformations and their rankings in the individual energy screens. Additional conformations are shown where all the bottom 10 are either high or low RMSD, to allow comparison of the individual component rankings betweem high and low RMSD structures.
Energies are in kcal/mole; RMSD is in angstroms.

1cgs; 8883 conformations

Conf	*Energy*	*RMSD*	*vdW*	*Ranking in Electrostatic*	*Internals*
4129	2414	2.4	53	4978	14
4375	2415	3.5	130	1280	1786
2952	2419	4.0	1136	8	4853
4402	2419	3.6	1229	592	1283
4383	2420	3.7	106	1253	2637
8690	2420	3.9	10	4773	540
2951	2421	4.0	1194	9	4953
4378	2421	3.4	580	1490	1251
4410	2421	3.5	692	365	2937
1483	2421	3.1	665	282	3354
(8163)	-	1.8	246	3798	290
(1211)	-	1.7	1030	1916	515
(8162)	-	1.7	110	15218	147
(8337)	-	1.9	223	5801	34

1mam; 5278 conformations

3317	3099	4.4	1	4226	5
3325	3102	4.3	6	3724	7
3316	3105	4.4	35	2958	12
3315	3105	4.4	17	3551	3
3319	3106	4.5	8	4276	1
3324	3108	4.3	12	3835	11
3318	3113	4.3	56	3783	4
3173	3113	1.8	549	1006	10
3207	3114	2.1	447	558	40
3048	3114	3.0	199	501	398
(3162)	-	1.8	544	876	16
(3196)	-	2.1	481	760	32
(3149)	-	2.0	625	748	22

1vfa; 4804 conformations

Conf	*Energy*	*RMSD*	*vdW*	*Ranking in Electrostatic*	*Internals*
1779	2079	1.8	26	371	44
1781	2082	1.6	43	400	37
1755	2083	1.6	10	965	7
1754	2085	1.6	15	936	12
1780	2085	1.7	54	493	23
4569	2088	2.2	6	716	197
4599	2089	2.6	5	1019	134
4452	2091	1.6	1	1540	230
4585	2093	2.3	13	1006	174
4568	2095	2.4	38	887	113

2fbj; 8082 conformations

6173	2835	1.9	4	2624	6
4424	2835	1.9	6	1167	25
6167	2836	1.9	14	1571	2
821	2837	2.4	17	2669	9
6171	2837	1.8	13	1701	14
6172	2837	1.8	21	1150	10
49	2839	1.9	2	5107	11
6168	2840	1.9	35	797	3
823	2841	2.4	9	1704	62
820	2842	2.4	27	959	21

1for; 6855 conformations

5864	2082	1.5	57	883	2
5869	2083	1.5	47	1171	1
5880	2083	1.4	51	1020	3
5877	2084	1.4	77	890	4
677	2085	1.9	1	853	354
5810	2087	1.6	58	625	65
5808	2089	1.6	88	559	70
5770	2090	1.8	78	1206	22
5874	2090	1.5	168	1004	5
5766	2091	1.8	101	968	35
(648)	2098	2.6	28	1762	331
(655)	2106	2.3	42	1151	735
(131)	2110	4.3	151	11	1827
(234)	2111	4.3	114	131	1982

1igf; 2431 conformations

Conf	*Energy*	*RMSD*	*vdW*	*Ranking in Electrostatic*	*Internals*
2028	2692	2.0	2	798	1
2039	2695	1.9	3	736	3
2037	2698	1.8	5	753	2
1989	2700	2.6	4	124	44
2001	2701	2.5	6	175	22
1990	2707	2.6	7	227	45
1999	2709	2.4	1	1123	29
2044	2712	1.8	14	792	9
2040	2714	1.9	31	117	7
2008	2716	2.3	20	146	48

1igm; 15815 conformations

15802	2602	2.8	562	1006	31
15814	2603	2.7	455	896	121
15764	2603	2.8	681	843	16
15803	2604	2.8	512	963	88
15809	2604	2.7	474	547	260
15806	2604	2.6	422	661	316
15815	2605	2.7	539	655	170
15796	2605	2.7	582	805	108
15694	2607	2.9	745	1177	38
15700	2607	2.9	762	1439	11
(15559)	2657	1.9	2385	1644	2840
(15576)	2665	1.9	3424	854	3006
(4545)	2668	2.0	2323	1475	5683

Table 5. Comparison of rms spread of 200 lowest energy loops, when screening with various VFF terms. NM = No minimisation (otherwise, 45 rounds of minimisation are done); All = spread of all generated conformations (otherwise, the spread of the 200 lowest in energy by that screen); VDW = van der Waals; Int = Internal energies; Elstat = electrostatics.

1bbj

RMSD range start	*All NM*	*Full VFF NM*	*All*	*Full VFF*	*VDW only*	*Elstat only*	*Int only*	*Full VFF without elstat*
0.5	45	29	74	41	34	14	38	37
1.0	319	84	653	107	111	74	83	108
1.5	921	55	1114	40	47	73	62	43
2.0	896	28	701	12	8	32	17	12
2.5	555	3	441	0	0	0	0	0
3.0	304	0	79	0	0	0	0	0
3.5	23	0	2	0	0	0	0	0
4.0	1	0	0	0	0	0	0	0

1cgs

1.0	1	0	4	0	0	0	0	0
1.5	160	1	232	11	1	0	13	9
2.0	332	4	342	4	3	0	26	7
2.5	636	17	1029	31	44	22	26	51
3.0	2099	83	2346	81	63	75	82	82
3.5	1931	41	1829	46	58	62	13	35
4.0	1369	34	1338	13	21	30	35	16
4.5	1172	4	983	12	3	8	5	0
5.0	677	15	428	2	7	1	0	0
5.5	228	1	197	0	0	2	0	0
>6	265	0	155	0	0	0	0	0

1mam

1.0	15	0	35	0	1	1	0	0
1.5	491	0	623	41	0	18	46	9
2.0	828	0	882	39	1	59	32	7
2.5	1025	7	1085	20	55	77	7	4
3.0	907	3	847	33	25	32	1	18
3.5	597	16	600	10	7	3	1	7
4.0	749	36	837	39	73	10	63	92
4.5	623	137	354	18	39	0	49	63

1vfa

1.0	-	-	9	1	0	0	0	-
1.5	-	-	653	79	49	31	13	-
2.0	-	-	922	69	31	44	12	-
2.5	-	-	455	15	15	53	3	-
3.0	-	-	691	20	24	35	0	-
3.5	-	-	1105	8	34	16	2	-
4.0	-	-	820	6	38	21	162	-
4.5	-	-	103	2	9	0	8	-
5	-	-	46	0	0	0	0	-

2fbj

RMSD range start	*All NM*	*Full VFF NM*	*All*	*Full VFF*	*VDW only*	*Elstat only*	*Int only*	*Full VFF without elstat*
1.0	27	2	181	29	26	0	21	30
1.5	701	65	1001	96	79	35	104	90
2.0	687	71	1640	44	49	79	49	45
2.5	1127	26	1003	15	20	40	8	16
3.0	908	26	1032	16	25	18	17	19
3.5	870	10	853	0	1	0	1	0
4.0	1219	0	1473	0	0	19	0	0
4.5	1135	0	816	0	0	9	0	0
>5.0	308	0	899	0	0	0	0	0

1igf

1.0	9	5	10	2	1	0	7	4
1.5	239	31	233	25	25	12	30	29
2.0	415	37	418	69	76	31	52	62
2.5	327	44	379	51	49	43	58	50
3.0	347	56	294	39	30	5	38	37
3.5	205	14	201	10	10	19	9	11
4.0	404	5	381	0	1	43	3	0
4.5	187	0	200	0	0	15	0	0
5.0	206	0	217	0	0	26	1	0
5.5	71	0	73	1	1	5	1	2
6.0	19	7	23	3	7	1	1	5
>6.5	2	1	2	0	0	0	0	0

1for

1.0	39	15	64	32	3	0	36	23
1.5	734	106	919	113	95	0	101	115
2.0	1767	64	1834	40	65	0	36	53
2.5	1026	15	906	6	23	16	27	8
3.0	587	0	568	0	0	0	0	0
3.5	575	0	689	0	0	0	0	0
4.0	782	0	773	9	10	101	0	1
4.5	817	0	774	0	4	79	0	1
5.0	475	0	304	0	0	4	0	0
>5.5	53	0	0	0	0	0	0	0

1hil

1.0	-	-	45	0	0	0	10	-
1.5	-	-	655	0	0	0	152	-
2.0	-	-	1194	9	5	16	23	-
2.5	-	-	3130	47	3	154	1	-
3.0	-	-	3877	92	175	136	2	-
3.5	-	-	3752	51	136	40	49	-
4.0	-	-	2410	284	162	128	121	-
4.5	-	-	1847	17	4	24	6	-
5.0	-	-	2683	0	9	1	72	-
5.5	-	-	1413	0	6	0	64	-
6.0	-	-	350	0	0	1	0	-

1igm

RMSD range start	*All NM*	*Full VFF NM*	*All*	*Full VFF*	*VDW only*	*Elstat only*	*Int only*	*Full VFF without elstat*
1.5	-	-	6	0	-	-	-	0
2.0	-	-	1148	0	-	-	-	0
2.5	-	-	1121	98	-	-	-	37
3.0	-	-	1249	41	-	-	-	35
3.5	-	-	3018	17	-	-	-	65
4.0	-	-	2970	39	-	-	-	55
4.5	-	-	1929	5	-	-	-	7
>5.0	-	-	4374	0	-	-	-	0

2.4 Conclusions

The unmodified AbM was not consistent in producing good models, or a spread of low RMSD conformations in the bottom 5 energies, with either the full or solvent-modified VFF. In fact, the conformations in the bottom 5 were generally the same, though in a slightly different order, whether the full or solvent-modified forcefield was used. This leads to the conclusion that the forcefield was not a significant factor in determining whether the lowest energy loops were selected.

Three possible solutions to the problem were explored. First, the CONGEN rebuild range was altered to produce more variation between models in the antigen-binding region; second, minimisation was used to attempt to releave strained but otherwise good structures, or to improve good structures with steric clashes. Third, individual components of the VFF force field were used to attempt to isolate the origin of the poor energy-RMSD correlations.

The first modification led to no improvement in the RMSD of the selected loops in one case, and a deterioration in the three others. In three of the four cases, the percentage of loops with RMSD below 2.0Å from the entire sample (Table 6) was examined for the

modelling runs using the original and altered rebuild range.

Table 6. The percentages of all conformations below 2Å for the original and altered rebuild range (modified VFF).

Structure	*Original*	*Altered*
1mam	12.9	7.83
1igf	10.6	6.83
2fbj	14.6	33.1

For two of these three, the percentage was around 5% lower for the altered range, so the more accurate conformations were being rejected at the CONGEN stage, probably due to bad contacts. In the other, 2fbj, it was around 20% higher, indicating that in this case a greater number of accurate conformations were being produced by CONGEN, but these were rejected at the VFF stage due to high energy.

Minimisation led to an improvement in RMSD distribution in most cases, although this was only noticeable if the bottom 200 conformations were considered. This supports the idea that a number of good conformations had high energies due to the framework/loop joining procedure, or were strained structures in general. It suggests that energy can be used as a screen to reduce the number of conformations to around 200, but to obtain further improvement, more discriminating screens are needed to identify the low RMSD loops.

Finally, in conjunction with the minimisation, the individual energy components of the VFF were used as a screen. It was shown that there is no consistent correlation between RMSD and any of the individual energy components. Indeed, for electrostatic energies, more high RMSD loops were selected in the bottom 200. Examination of structures with Insight II showed that this was due to formation of non-native salt bridges, particularly in 1hil which has a large number of charged residues. In view of this, a VFF screen containing all but the electrostatics terms was tried for selected structures (Table 5). However, no improvement in RMSD/energy correlation was seen, illustrating that other factors (such as repulsive van der Waals) may also contribute to the selection of high RMSD structures.

When the low RMSD/high energy conformations were examined by PROCHECK (Laskowski et al, 1993), they were found to be poorer quality structures, with more Ramachandran disallowed torsions (Table 6), and in some cases having bond lengths and angles that deviated from the ideal.

In summary, despite some considerable changes to the old AbM, there remained some problems with accurate modelling of H3 loops. Two main causes suggested themselves:

a) Incomplete energy force field (other terms need to be accounted for, such as solvation)
b) The possibility of potentially 'good' loops having high energy due to localised interactions, such as sidechain clashes, that are not completely removed by minimisation.

These possibilities will be examined in subsequent chapters.

CHAPTER 3 - The roles of solvation and electrostatics in CDR modelling

3.1 Background

An important factor in protein surface loop conformation is solvation by water. The presence of water will stabilise to some extent exposed, unpaired charged residues which would not otherwise be favoured, and will disfavour exposed non-polar residues due to the hydrophobic effect. In its 'old' version, as seen in the previous Chapter, AbM attempts to account for the solvent by turning off the attractive terms in the VFF force field. Since this has been shown to be ineffectual, other methods are needed.

There are a number of approaches in the treatment of solvation effects. For example, a molecular dynamics simulation with explicit water can be carried out, in which the loop is surrounded by water molecules and allowed to stabilise into a low energy form (McCammon and Harvey, 1987). This method has the advantage that it most closely simulates the actual situation (assuming the force field used accurately describes all the appropriate energies) but has the disadvantage that it is computationally intensive and, if a large number of conformations are to be examined, is not the method of choice.

Other approaches are based on simple accessibility algorithms or continuum electrostatic calculations. Typical of the former is the method of Eisenberg and McLachlan (1986). This method consists of multiplying the accessible surface area of each atom (α) with the hydrophilicity of that atom, expressed as an **atomic solvation parameter** (ASP,σ). Atoms are divided into five types of varying hydrophilicity; charged N, charged O, neutral N or O, C or S. The overall energy is the sum of the energies for each atom:

$$E = \Sigma_{N+} \alpha_{N+}\sigma_{N+} + \Sigma_{O-} \alpha_{O-}\sigma_{O-} + \Sigma_{N/O} \alpha_{N/O}\sigma_{N/O} + \Sigma_{C} \alpha_{C}\sigma_{C} + \Sigma_{S} \alpha_{S}\sigma_{S}$$

The actual parameters were derived from transfer energies of amino acid analogues from octanol to water (Fauchere and Pliska, 1983) and are as follows.

Atom type	σ (kcal A^{-2} mol^{-1})
Charged N	-0.050 +/- 0.009
Charged O	-0.024 +/- 0.010
Neutral N/O	-0.006 +/- 0.004
C	0.016 +/- 0.002
S	0.021 +/- 0.010

As can be seen, hydrophilic atoms have negative σ values and hydrophobic atoms positive σ values. As the equation shows, molecules that expose polar areas and bury hydrophobic areas have lower energy, whereas molecules that expose hydrophobic areas and bury polar areas have higher energy.

Another accessibility-based method is that of Kurochkina and Lee (1995). This is somewhat less comprehensive as it models the hydrophobic effect only - it does not reward exposed charged residues. The **pairwise surface area sum** is defined as the sum of the two areas that two atoms of a pair bury in contact, and is as below:

$$a_{ij} = 2\pi (R + R_p) (2R + 2R_p - d_{ij})$$

where a_{ij} is the pairwise surface area sum, R is the sum of the van der Waals radii of the two atoms, R_p is the effective radius of a water molecule and d_{ij} is the interatomic distance.

The hydrophobic energy is assumed to be proportional to the pairwise surface area sum of each pair, summed over all atom pairs, i.e.

$$E = k.\Sigma_{\text{all pairs}}\, a_{ij}$$

There are a number of problems with this method. Firstly, taking the pairwise sums over all pairs will lead to overcounting of atoms, as

atoms have more than one neighbour. However, if pairs involving an atom i and an atom j which is either in the same or adjacent residues to that of atom i are discounted, the overcounting is much reduced, and the pairwise sum energy correlates with the Fauchere and Pliska (1983) octanol-water transfer energy. The other main problem is that all atoms are treated as equal in hydrophobicity (there are no atomic solvation parameters as for Eisenberg and McLachlan) which is not the case in reality. This is dealt with by only including carbon atoms. However, it does seem a serious omission not to take explicitly into account charged and polar atoms, as without them, the preference for a lysine sidechain for example, with four carbon atoms, would alter from being exposed to being buried.

The continuum electrostatic approaches treat the solvent as a homogeneous area of high dielectric constant, and calculate the solvation energy by a number of electrostatics-based methods, taking this solvent property into account. The initial model was the Born model, (Born, 1920) which calculates the energy of introducing a charged sphere to water by summing the energy of discharging a sphere in vacuum and charging a sphere in water (it is assumed that introducing a neutral sphere into water uses no energy). The method has the disadvantages of ignoring the dipole interactions formed by the water on introduction of the sphere, and the cavitation (cavity-forming) energy. Most importantly, it cannot be used with non-spherical molecules.

The continuum approach has been extended in a number of ways, for example by the use of distance-dependent dielectric constants, which account for the solvent by increasing the dielectric constant depending on the distance between two charges (Gelin and Karplus, 1975; Weiner et al, 1984) and reducing the charge on atoms with increasing accessibility (Northrup et al, 1981).

Another approach has been taken by the group of Honig (Klapper et al., 1986; Gilson and Honig, 1988; Smith and Honig, 1994), who implement an algorithm to solve the Poisson-Boltzmann equation (Klapper et al.,1986) within the program DelPhi. The equation is solved (see Appendix 4) using a finite-difference method, in which the molecule is placed within a grid, with the atomic charges placed on the nearest grid corners to the given atom, and the corresponding potentials calculated at the grid corners. This is a limitation since often charges will be displaced from the grid positions so that the calculated electrostatic energy will not be an accurate reflection of the real distribution but it considerably simplifies the calculation.

To calculate the solvation energy (Gilson and Honig, 1988) the potential at each grid point is calculated with the protein in two different environments: first, in a medium with the dielectric constant

of the protein (dielectric 2), and second, in water (dielectric 80). The solvation energy is then given by:

$$dG_{solv} = 1/2\Sigma_i(q_i.\Delta\phi_i)$$

where q_i is the charge of grid point i, and $\Delta\phi_i$ is the difference in potentials in the two different environments at i. This is essentially equivalent to calculating the electrostatic energy in the two different environments using the effective dielectric constant at each atom, which depends upon its accessibility.

DelPhi can be used as one component in an algorithm to evaluate the total solvation energy, as shown below (Sitkoff et al, 1994):

$$E_{solv} = E_{vdW,protein\text{-}water} + E_{electrostatic} + E_{cavitation}$$

Indeed, the solvation can be combined with a standard energy calculation program such as VFF.

$$E_{tot} = E_{internals} + E_{vdW,intramolecular} + E_{vdW,protein\text{-}water} + E_{electrostatic} + E_{cavitation}$$

The van der Waals interactions between the protein and the water, and the cavitation energy, which is the energy required to form and maintain a protein-shaped cavity in the water, are accounted for in one

energy evaluation, based on the accessible surface area.

$$E_{cav/vdW} = \gamma A + b$$

where A is the exposed surface area ($Å^2$), and γ and b are constants obtained from alkane vacuum/water transfer free energies. γ has the value 0.005 kcal/mol/$Å^2$ and b has the value 0.00086 kcal/mol. Sitkoff et al (1994) developed a charge parameter set, PARSE, for use here which agrees considerably better than other parameter sets with the experimental solvation energies of a range of small organic molecules.

3.2 Methods

Eisenberg and McLachlan

The accessible surface areas of every atom in all the CAMAL-generated minimised H3 conformation of each modelled Fv structure were calculated, using the algorithm in DSSP (Kabsch and Sander, 1983). Each value was multiplied by the appropriate ASP value to give an energy for that atom, and the energies for each atom were summed to give an overall energy. The bottom 10 conformations were then ranked according to their energy, and the RMSD spread of the bottom 200 conformations was noted.

Additionally, the 200 lowest energy conformations by VFF were screened (Figure 14) using this method, and the bottom 10 conformations ranked; this enabled the non-solvation terms to be taken into account.

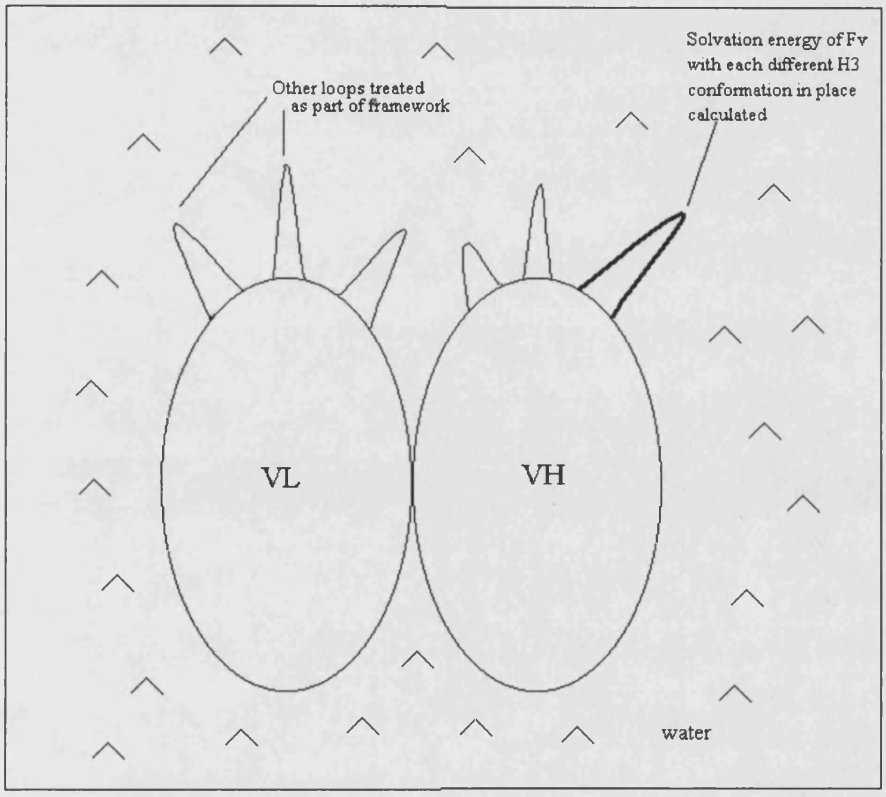


Figure 14. Solvation energy calculation. The solvation energy of the Fv with each H3 conformation in place is calculated; the conformation of the H3 will affect not only the accessibility, and therefore energy, of H3 residues but may also affect that of the other CDRs.

DelPhi

The minimised CAMAL-generated H3 loop conformations for each minimised structure were subjected to torsional clustering (Appendix 2), with an initial resolution and step value of 15 degrees and a target of 1000 conformations. This was necessary due to the time taken by DelPhi (5 minutes per conformation) which would have made unreasonable time demands with up to 8000 conformations for some structures.

Each clustered conformation was patched onto the entire Fv. The grid size was calculated such that the entire Fv would fit into the grid with a fill of 90%, and such that the grid resolution was 2 grid points/Å (this resolution maximises accuracy - see introduction - while not making unreasonable time demands). The electrostatic energy calculated with DelPhi using both an external dielectric of 2, and an external dielectric of 80, representing desolvated and solvated states respectively (the interior dielectric was 2 in both cases). The electrostatic component to the solvation energy was calculated by taking the difference in the electrostatic energies obtained in the output file for each run.

The conformations were ranked both using DelPhi alone, and using the combined VFF/DelPhi/protein-water van der Waals and cavitation screen (see above). DelPhi was used with a grid resolution

of 2 grids/Å and a molecule grid fill of 90% (DelPhi manual). Sitkoff et al. (1984) obtain the best results using the PARSE parameter set; however, in general, VFF parameters for atomic charges were used here, for consistency, with a couple of exceptions (see below).

In another approach to combining DelPhi and VFF, the 200 lowest energy VFF conformations (from the entire set, not the clustered set) were screened using DelPhi. For two of the structures (1mam and 1vfa), the PARSE parameter set was used in addition to the VFF set, in order to compare the two.

3.3 Results

Eisenberg and McLachlan

Tables 7 and 7a show that the Eisenberg and McLachlan method is generally worse than VFF at selecting low RMSD conformations. When using Eisenberg and McLachlan only, 3 out of 4 structures give a higher RMSD distribution of conformations in the lowest 10 in energy compared with VFF, and the fourth, 1cgs, gives a similar distribution (it remains poor with no loops < 2Å selected). This pattern is consistent when the RMSD distribution of the bottom 200 loops is examined (Figure 15), with the possible exception of 1igf. In this case, more conformations of RMSD < 2Å are selected, but also more conformatons of RMSD > 3.5Å are selected.

The use of Eisenberg and McLachlan as a screen for the lowest 200 VFF conformations is no more encouraging. In 3/5 structures, the bottom 10 conformations by Eisenberg and McLachlan have a higher RMSD distribution than the bottom 10 by VFF. An improvement is seen in 1mam, with the conformation of RMSD < 2Å improving in position from 8th to 3rd, and 1hil, where 6/10 conformations have RMSD < 3Å with the solvation method, compared to 10/10 with RMSD >= 4Å with VFF.

Table 7. Comparison of the RMSD spread of VFF the Eisenberg and McLachlan screen used by itself and used as a screen for the bottom 200 VFF conformations.
Energies are in kcal/mole; RMSD is in angstroms.

1bbj

VFF			Eisenberg			Eisenberg as screen		
Conf	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*
2544	2661	1.1	-	-	-	475	-3.963	1.5
931	2661	1.5	-	-	-	2848	-3.013	1.5
1198	2662	1.2	-	-	-	527	-2.813	1.1
2151	2662	0.6	-	-	-	2555	-2.758	0.7
389	2662	0.4	-	-	-	2235	-2.637	2.4
1781	2663	0.5	-	-	-	666	-2.602	1.8
2836	2663	0.9	-	-	-	2151	-2.389	0.6
472	2664	0.9	-	-	-	1781	-2.361	0.5
1745	2665	1.1	-	-	-	330	-2.254	1.4
1647	2666	1.9	-	-	-	1717	-2.132	2.0
1cgs								
4129	2414	2.4	2999	-6.191	3.3	3403	-3.933	3.3
4375	2415	3.5	943	-6.093	3.2	6266	-3.443	2.9
2952	2419	4.0	959	-6.001	2.9	6267	-3.326	2.8
4402	2419	3.6	3000	-6.001	3.2	6268	-3.258	2.8
4383	2420	3.7	934	-5.987	3.1	6270	-3.219	2.8
8690	2420	3.9	944	-5.973	2.9	4413	-3.036	3.6
2951	2421	4.0	2358	-5.966	2.3	3391	-2.573	3.1
4378	2421	3.4	945	-5.904	2.9	3408	-2.560	3.0
4410	2421	3.5	4448	-5.880	3.1	4393	-2.514	3.4
1483	2421	3.1	953	-5.857	2.9	1493	-2.491	3.2
1mam								
3317	3099	4.4	-	-	-	1009	-2.707	2.8
3325	3102	4.3	-	-	-	5007	-2.689	2.3
3316	3105	4.4	-	-	-	3268	-2.406	1.8
3315	3105	4.4	-	-	-	1026	-1.720	2.6
3319	3106	4.5	-	-	-	4548	-1.354	3.1
3324	3108	4.3	-	-	-	4547	-1.334	3.1
3318	3113	4.3	-	-	-	5027	-1.227	2.4
3173	3113	1.8	-	-	-	5016	-1.149	2.0
3207	3114	2.1	-	-	-	1017	-1.137	2.7
3048	3114	3.0	-	-	-	1019	-1.090	2.6

1vfa

VFF			Eisenberg			Eisenberg as screen		
Conf	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*
1779	2079	1.8	-	-	-	384	-8.256	2.4
1781	2082	1.6	-	-	-	387	-8.066	2.2
1755	2083	1.6	-	-	-	377	-7.218	2.3
1754	2085	1.6	-	-	-	373	-7.216	2.5
1780	2085	1.7	-	-	-	374	-6.435	2.5
4569	2088	2.2	-	-	-	410	-6.293	2.1
4599	2089	2.6	-	-	-	3264	-5.582	2.1
4452	2091	1.6	-	-	-	1770	-5.271	1.5
4585	2093	2.3	-	-	-	395	-5.128	2.0
4568	2095	2.4	-	-	-	380	-4.901	2.2
2fbj								
6173	2835	1.9	4985	-4.568	4.7	-	-	-
4424	2835	1.9	751	-3.924	2.0	-	-	-
6167	2836	1.9	912	-3.718	4.8	-	-	-
821	2837	2.4	776	-3.631	2.2	-	-	-
6171	2837	1.8	6899	-3.604	1.3	-	-	-
6172	2837	1.8	7750	-3.583	4.8	-	-	-
49	2839	1.9	910	-3.511	4.7	-	-	-
6168	2840	1.9	748	-3.504	1.9	-	-	-
823	2841	2.4	6904	-3.478	1.1	-	-	-
820	2842	2.4	760	-3.460	2.2	-	-	-
1igf								
2028	2692	2.0	92	-5.473	2.7	-	-	-
2039	2695	1.9	97	-5.320	2.7	-	-	-
2037	2698	1.8	1798	-5.131	3.3	-	-	-
1989	2700	2.6	1807	-4.892	3.3	-	-	-
2001	2701	2.5	1788	-4.794	3.3	-	-	-
1990	2707	2.6	101	-4.612	2.7	-	-	-
1999	2709	2.4	100	-4.055	2.8	-	-	-
2044	2712	1.8	1793	-3.773	3.4	-	-	-
2040	2714	1.9	1781	-3.684	3.5	-	-	-
2008	2716	2.3	1799	-3.676	3.4	-	-	-

1for

VFF			Eisenberg			Eisenberg as screen		
Conf	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*
5864	2082	1.5	6752	-2.283	2.4	6344	-0.433	2.3
5869	2083	1.5	6757	-2.117	2.3	6346	-0.266	2.2
5880	2083	1.4	6807	-2.112	2.4	6345	-0.085	2.2
5877	2084	1.4	6800	-2.045	2.4	6338	0.116	2.2
677	2085	1.9	6804	-2.017	2.4	671	0.656	2.0
5810	2087	1.6	6803	-2.016	2.4	5903	0.736	1.4
5808	2089	1.6	6808	-1.973	2.4	684	0.879	1.7
5770	2090	1.8	6797	-1.959	2.5	1048	1.012	1.8
5874	2090	1.5	6801	-1.951	2.4	1033	1.127	1.9
5766	2091	1.8	6791	-1.857	2.3	691	1.187	1.8
1hil								
3362	2093	4.1	-	-	-	6832	-4.573	2.6
3557	2093	4.3	-	-	-	6677	-4.532	2.5
3590	2094	4.1	-	-	-	6831	-4.453	2.6
3395	2094	4.3	-	-	-	2335	-3.768	4.0
3164	2097	4.0	-	-	-	2135	-3.467	4.1
3416	2099	4.2	-	-	-	6913	-3.293	2.5
3357	2099	4.1	-	-	-	2217	-3.128	4.0
3579	2099	4.3	-	-	-	6531	-3.077	2.6
3366	2101	4.0	-	-	-	2268	-2.782	4.0
3367	2101	4.0	-	-	-	6515	-2.284	2.6

TABLE 7a. The RMSD spread of the 200 lowest Eisenberg energy conformations, and the VFF spread for comparison.

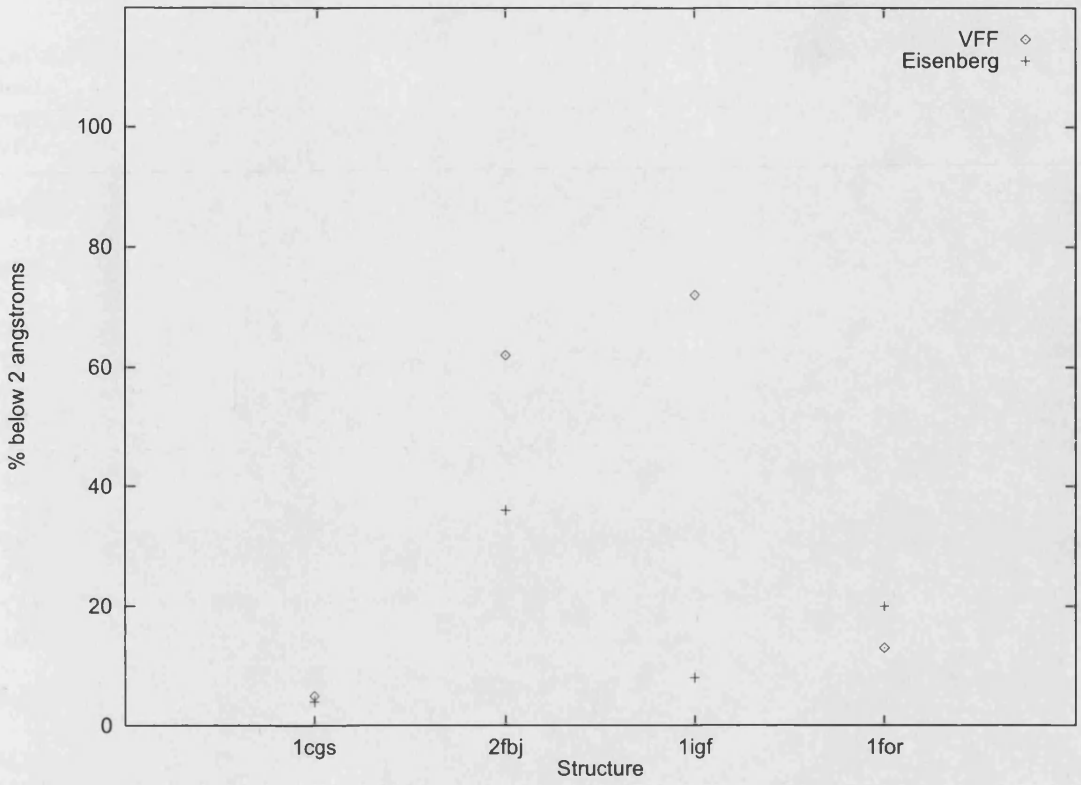
RMSD range	1cgs		2fbj		1for		1igf	
start	Eis	VFF	Eis	VFF	Eis	VFF	Eis	VFF
1	3	55	5	4				
1.5	5	19	68	1	16	11	41	
2	3	63	42	68	74	110	8	39
2.5	50	74	5	26	40	60	49	53
3	118	29	15	25	6	14	87	44
3.5	14	18	31	8	12	43	9	
4	12	4	14	6	2	1		
4.5	3	4	19	61				
5								
5.5	3							
6	6							

Using the combined and DelPhi screen on all conformations

As seen in Tables 8 and 8a, neither the combined nor the DelPhi only screens showed an improvement on the performance of VFF.

Using the combined screen, the 10 lowest energy conformations show a similar RMSD distribution to VFF in all cases. Worthy of note, however, is that one loop with RMSD less than 2Å is appearing in the 10 selected conformations in 1cgs: this was not observed in VFF. This would appear to be an isolated case, however, as many of the other loops in the 10 have RMSD approaching or exceeding 5Å, and the RMSD spreads of the lowest energy 200 conformations for all 4 structures are rather less good than with VFF (Figure 15).

When DelPhi only is used as a screen, the results in general deteriorate with respect to either the VFF or the combined screen, with higher (considerably for 2fbj) RMSD loops in both the 10 and the 200 lowest energy conformations (Figure 15). An exception to this behaviour is seen in 1cgs, in which 3 of the 10 lowest energy conformations have RMSD below 2Å, and a further 2 not much above 2Å.. This is a considerable improvement over VFF. The 200 lowest energy loops for 1cgs (Table 8a) show a somewhat different pattern. Although there are more conformations in the 1.5 - 2.5Å RMSD range compared to VFF, the peak density of conformations is found in the 3.0-5.0Å RMSD range, compared to 2.5 - 4.0Å with VFF. The



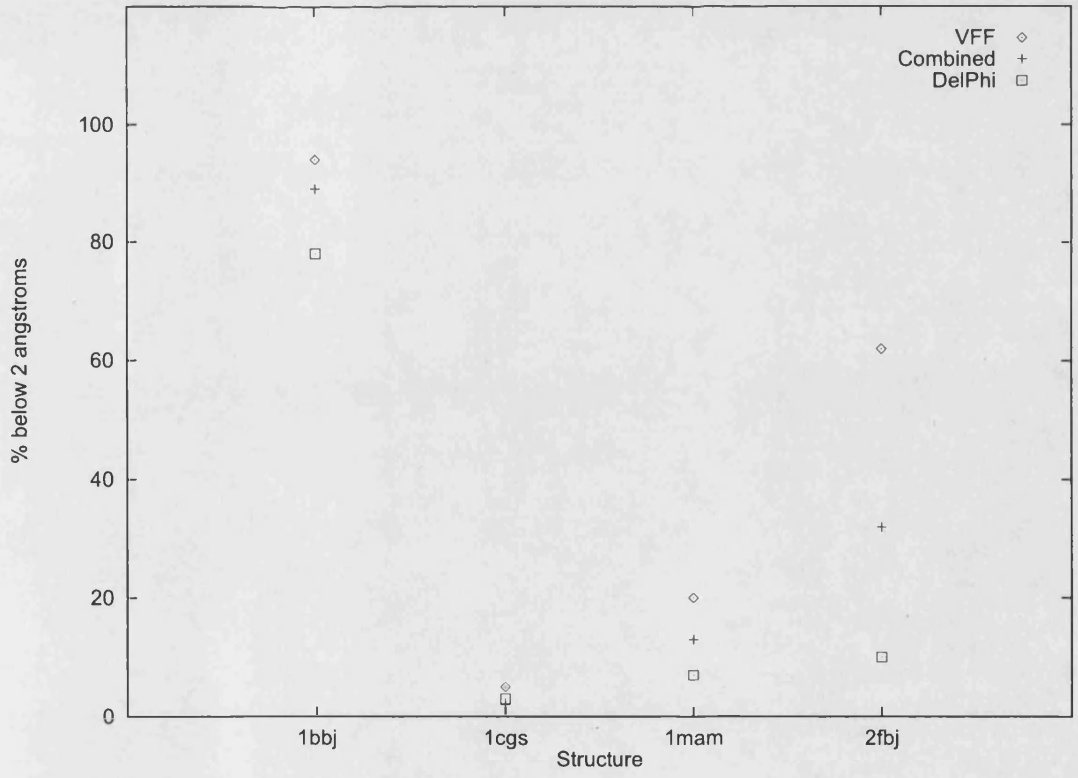


Figure 15. The percentage of the bottom 200 energy H3 conformations below 2 angstroms: VFF vs. Eisenberg (above), VFF vs. Combined DelPhi/Cavitation/VFF vs. DelPhi (below).

Table 8. Comparison of the RMSD spread of VFF, DelPhi used by itself **(Del),** DelPhi combined with VFF and cavitation/protein-water van der Waals **(Cmb),** and DelPhi as a screen for the bottom 200 VFF conformations **(Scrn).**
VFF charges are used in DelPhi; energies are kcal/mole and RMSD is in angstroms.

1bbj

VFF			*Del*			Cmb			*Scrn*		
Conf	Energy	RMSD	*Conf*	*Energy*	*RMSD*	Conf	Energy	RMSD	*Conf*	*Energy*	*RMSD*
2544	2661	1.1	*2544*	*1602*	*1.1*	2390	-1169	1.5	*1778*	*-1169*	*0.8*
931	2661	1.5	*1645*	*1609*	*0.8*	1407	-1169	1.3	*1986*	*-1169*	*1.1*
1198	2662	1.2	*389*	*1614*	*0.4*	981	-1168	1.7	*576*	*-1168*	*1.4*
2151	2662	0.6	*1723*	*1619*	*1.0*	1986	-1168	1.1	*1636*	*-1167*	*1.2*
389	2662	0.4	*2635*	*1619*	*1.0*	576	-1168	1.4	*2433*	*-1167*	*1.0*
1781	2663	0.5	*2766*	*1619*	*0.6*	1280	-1168	1.6	*2177*	*-1166*	*0.9*
2836	2663	0.9	*2850*	*1621*	*1.3*	870	-1167	1.5	*1080*	*-1166*	*0.9*
472	2664	0.9	*1271*	*1623*	*2.0*	2869	-1167	1.5	*1645*	*-1166*	*0.8*
1745	2665	1.1	*576*	*1626*	*1.4*	370	-1167	1.5	*2555*	*-1165*	*0.7*
1647	2666	1.9	*3063*	*1626*	*1.4*	273	-1167	1.2	*2893*	*-1165*	*1.1*

1cgs

4129	2414	2.4	*5495*	*1237*	*4.9*	4805	-1302	2.8	*1763*	*-1261*	*3.0*
4375	2415	3.5	*735*	*1238*	*4.7*	1398	-1302	1.5	*2608*	*-1260*	*3.3*
2952	2419	4.0	*7834*	*1239*	*2.5*	2156	-1294	1.6	*5394*	*-1260*	*4.1*
4402	2419	3.6	*1208*	*1239*	*1.7*	2155	-1293	2.1	*2610*	*-1259*	*3.3*
4383	2420	3.7	*1219*	*1239*	*5.1*	7676	-1292	2.3	*2616*	*-1258*	*3.4*
8690	2420	3.9	*754*	*1241*	*4.8*	4032	-1289	3.0	*2614*	*-1258*	*3.4*
2951	2421	4.0	*1960*	*1242*	*3.1*	2160	-1288	1.7	*2617*	*-1257*	*3.2*
4378	2421	3.4	*5529*	*1244*	*4.8*	8124	-1288	4.2	*694*	*-1257*	*4.6*
4410	2421	3.5	*8143*	*1244*	*4.7*	7032	-1287	3.8	*2615*	*-1257*	*3.3*
1483	2421	3.1	*405*	*1245*	*3.4*	6762	-1286	3.5	*5562*	*-1255*	*5.1*

1mam

3317	3099	4.4	*3325*	*2208*	*4.3*	3618	-968	3.8	*3601*	*-961*	*3.6*
3325	3102	4.3	*3316*	*2215*	*4.4*	3747	-965	4.1	*3474*	*-961*	*3.6*
3316	3105	4.4	*3321*	*2217*	*4.5*	1617	-964	4.0	*3577*	*-959*	*4.5*
3315	3105	4.4	*88*	*2219*	*4.3*	1307	-964	4.1	*3562*	*-959*	*4.6*
3319	3106	4.5	*3320*	*2220*	*4.4*	1218	-962	4.0	*3590*	*-958*	*4.4*
3324	3108	4.3	*3707*	*2224*	*4.5*	3378	-962	4.0	*3543*	*-957*	*4.7*
3318	3113	4.3	*3314*	*2226*	*4.3*	3755	-962	4.3	*3589*	*-957*	*4.4*
3173	3113	1.8	*3562*	*2227*	*4.6*	3596	-961	3.6	*3713*	*-957*	*4.4*
3207	3114	2.1	*1947*	*2228*	*4.8*	3694	-961	4.0	*3707*	*-957*	*4.5*
3048	3114	3.0	*86*	*2229*	*4.2*	1139	-961	3.3	*3579*	*-957*	*4.5*

1vfa

VFF			*Del*			Cmb			*Scrn*		
Conf	Energy	RMSD	*Conf*	*Energy*	*RMSD*	Conf	Energy	RMSD	*Conf*	*Energy*	*RMSD*
1779	2079	1.8	-	-	-	-	-	-	*387*	*-999*	*2.2*
1781	2082	1.6	-	-	-	-	-	-	*373*	*-998*	*2.5*
1755	2083	1.6	-	-	-	-	-	-	*384*	*-997*	*2.4*
1754	2085	1.6	-	-	-	-	-	-	*3202*	*-992*	*3.8*
1780	2085	1.7	-	-	-	-	-	-	*377*	*-992*	*2.3*
4569	2088	2.2	-	-	-	-	-	-	*410*	*-991*	*2.1*
4599	2089	2.6	-	-	-	-	-	-	*374*	*-991*	*2.5*
4452	2091	1.6	-	-	-	-	-	-	*1767*	*-990*	*1.6*
4585	2093	2.3	-	-	-	-	-	-	*3307*	*-989*	*3.2*
4568	2095	2.4	-	-	-	-	-	-	*378*	*-989*	*2.4*
2fbj											
6173	2835	1.9	*6173*	*1797*	*1.9*	5871	-1113	4.3	*818*	*-1101*	*2.3*
4424	2835	1.9	*819*	*1805*	*2.3*	5406	-1111	3.8	*819*	*-1099*	*2.3*
6167	2836	1.9	*7040*	*1811*	*1.8*	3042	-1108	3.1	*48*	*-1099*	*2.2*
821	2837	2.4	*6058*	*1812*	*3.3*	975	-1108	3.2	*85*	*-1098*	*2.0*
6171	2837	1.8	*4417*	*1813*	*2.1*	2434	-1108	4.4	*44*	*-1098*	*2.1*
6172	2837	1.8	*7041*	*1814*	*1.8*	852	-1108	4.8	*6915*	*-1097*	*1.2*
49	2839	1.9	*6212*	*1816*	*1.6*	2385	-1107	4.1	*5486*	*-1097*	*2.8*
6168	2840	1.9	*7011*	*1817*	*1.5*	5296	-1107	4.5	*6062*	*-1097*	*3.3*
823	2841	2.4	*7084*	*1819*	*2.1*	3935	-1107	4.1	*816*	*-1097*	*2.4*
820	2842	2.4	*60*	*1821*	*1.8*	5249	-1107	3.9	*86*	*-1097*	*2.0*

TABLE 8a. The RMSD spread of the 200 lowest energy conformations as selected by the combined VFF,DelPhi and cavitation/protein-water van der Waals screen, and DelPhi only, with the VFF spread for comparison.

	1bbj			1cgs		
RMSD range start	VFF	Combined	DelPhi	VFF	Combined	DelPhi
0.5	42	11	7	0	0	0
1.0	107	85	55	0	0	0
1.5	39	82	95	11	5	7
2.0	12	17	34	4	7	16
2.5	0	5	8	31	33	17
3.0	0	0	0	81	40	26
3.5	0	0	1	46	41	34
4.0	0	0	0	13	22	52
4.5	0	0	0	12	26	28
5.0	0	0	0	2	15	9
>5.5	0	0	0	0	11	15

	1mam			2fbj		
RMSD range start	VFF	Combined	DelPhi	VFF	Combined	DelPhi
1.0	0	3	1	29	9	1
1.5	41	24	13	96	56	20
2.0	39	21	10	44	62	27
2.5	20	33	18	15	30	15
3.0	33	35	28	16	22	20
3.5	10	15	37	0	16	18
4.0	39	43	71	0	5	63
4.5	18	24	21	0	0	32
5.0	0	2	1	0	0	4

improved results are unlikely to be meaningful; if DelPhi energy has no correlation with RMSD (as evidenced from the other structures), it is likely to occasionally pick good structures.

DelPhi as a screen for the lowest energy VFF loops

As shown in Table 8, using DelPhi to screen the 200 lowest VFF energy conformations is no more effective at selecting low RMSD conformations than DelPhi on its own, or the combined screen. Considering the 10 lowest energy loops, for 2 out of the 5 structures (1cgs and 1mam) poor loops are again selected, with a similar RMSD distribution to that of VFF. For the other 3 structures, 2fbj, 1igm and 1vfa, the selected conformations with the DelPhi screen generally have higher RMSD than those selected with VFF.

Lastly, substituting the VFF with the PARSE parameter set (Table 9) in 1mam and 1vfa does not improve the results: indeed in 1vfa, slightly higher RMSD loops are selected.

Table 9. Comparison of the VFF and PARSE charge parameter set when using DelPhi as a screen for the bottom 200 VFF conformations.

1mam

VFF			Parse		
Conf	*Energy*	*RMSD*	*Conf*	*Energy*	*RMSD*
3601	-961	3.6	3601	-1462	3.6
3474	-961	3.6	3474	-1461	3.6
3577	-959	4.5	3577	-1451	4.5
3562	-959	4.6	3562	-1450	4.6
3590	-958	4.4	3589	-1450	4.4
3543	-957	4.7	3590	-1449	4.4
3589	-957	4.4	3707	-1449	4.5
3713	-957	4.4	3713	-1449	4.4
3707	-957	4.5	3588	-1448	4.4
3579	-957	4.5	3579	-1448	4.5

1vfa

387	-999	2.2	3307	-1497	3.2
373	-998	2.5	387	-1492	2.2
384	-997	2.4	384	-1491	2.4
3202	-992	3.8	373	-1491	2.5
377	-992	2.3	1757	-1487	1.9
410	-991	2.1	3202	-1484	3.8
374	-991	2.5	3339	-1484	3.5
1767	-990	1.6	1764	-1483	2.0
3307	-989	3.2	3328	-1483	3.6
378	-989	2.4	3210	-1483	4.0

3.4 Conclusions

Eisenberg and McLachlan

The Eisenberg potential is not able to select low RMSD loops, whether from the entire set or from the 200 lowest energy VFF conformations, and does not rate the crystal structures with particularly low energy. Although the crystal structure might be expected to bury hydrophobic and expose charged groups, in practice this does not always happen. Examination of various crystal structure loops shows some examples of exposed hydrophobic (particularly aromatic) groups, which may be important for antigen binding, and buried charged groups forming salt bridges. Hence, conformations with low solvation energies may be unrealistic structures, and other energy terms may be more important. Nevertheless, solvation energy would be expected to play an important role in the conformational stability of all protein surface loops.

Again, when screening the 200 lowest VFF energy conformations, the lower RMSD conformations may not have such favourable solvation energy due to exposed aromatic sidechains and the formation of internal salt bridges. An improved solvation energy screen might therefore include a term to score favourably internal salt

bridges, and those residue types known to be exposed in antibody crystal structures.

DelPhi

Although some promising results were occasionally obtained with the DelPhi screen, they were not consistent enough to recommend DelPhi as a screening method in addition to, or as an alternative to, VFF. In fact, in the majority of examples DelPhi performed considerably worse than VFF, selecting high RMSD in preference to low RMSD conformations.

One potential problem with DelPhi is the fact that the charges of the molecule are represented on a grid, with the charges of each atom being placed on the nearest grid point. This means that the charges will be represented differently depending on the positioning of the molecule within the grid, leading to different results; this, combined with the fact that the centre of geometry of the molecule is placed in the centre of the grid, means that results for two different molecules cannot be compared. It also means that if a considerable part of a model of a given molecule has a different conformation to the reference molecule (for example the crystal structure), comparison of the results will not be valid. The high resolution of 2 grid points/Å should minimise this, as should the fact that here only the H3 loop is altered; however, it is still a potential source of error. A further

problem is that in the combined VFF/ DelPhi/ cavitation/ solvation van der Waals method, the energies may not be in the same scale relative to one another. A solution to this problem would involve introducing scaling factors for the various terms; the values would be those which give best agreement with experimental results for a range of examples. This should be returned to.

CHAPTER 4 - AN INVESTIGATION OF LOOP ENERGIES IN H3 SCREENING

4.1 Introduction

The purpose of this investigation was to determine why some of the low RMSD models have high energies relative to the lowest energy models. This was approached using progressive dissection of the energies; first the total energy can be broken down into the individual VFF terms such as van der Waals, internals and electrostatics, and then each energy component can be broken down still further to discover from where in the loop the high energy derives.

4.2 Examining the individual components

Individual terms of the VFF energy

For five structures, 7 high energy/low RMSD, and 7 low energy, minimised models were analysed. The low RMSD models were the lowest RMSD structures with overall energies at least 50kcal higher than the lowest energy structures. In addition the 7 structures were selected so that they belonged to at least four RMSD clusters (see Appendix 1). The low energy models were actually the 7 lowest

energy models, which belonged to at least four different RMSD clusters.

The individual VFF terms were noted and compared; they were also compared with those of the crystal structure - the crystal structures were checked to make sure they were low energy structures; this was confirmed, each having a lower energy than any of the models.

Table 10 shows that the predominant energy component contributing to the higher energy of the selected low rms models, doing so in 4 out of the 5 structures, is the repulsive van der Waals. In 1mam and 1vfa, the angle bending energy is also important (in 1mam more so than the repulsive van der Waals), and in 1cgs, and 1vfa to a slight extent, the electrostatic energy is important. In Chapter 2 it has been shown that charged sidechains can lead to formation of 'false' salt bridges.

Partitioning repulsive van der Waals

On the basis of the above results, it was decided to break down the repulsive van der Waals still further, to pinpoint the source of the high energies. For each conformation examined above, the energy was partitioned into intra-H3 energy and four classes of inter-H3-Fv (referred to as FR, although including the other 5 CDRs

Table 10. Dissection of the total energy of the high (bold) and low (italic) energy conformation sets into individual VFF components.

Conf	Total energy (kcal/mole)	Repulsive vdW	Dispersion vdW	Electro-static	Bond	Angle	Torsion
1cgs							
7145	**2657**	**497**	**-265**	**60**	**1168**	**881**	**282**
4071	**2505**	**351**	**-226**	**19**	**1156**	**872**	**295**
4072	**2507**	**352**	**-226**	**19**	**1156**	**873**	**298**
4070	**2514**	**351**	**-224**	**25**	**1156**	**871**	**299**
1318	**2487**	**341**	**-215**	**18**	**1161**	**872**	**276**
5284	**2526**	**383**	**-236**	**18**	**1160**	**888**	**279**
1397	**2751**	**644**	**-280**	**30**	**1166**	**881**	**276**
4129	*2414*	*290*	*-210*	*10*	*1159*	*866*	*271*
4375	*2415*	*286*	*-203*	*-12*	*1160*	*872*	*277*
2952	*2419*	*296*	*-197*	*-42*	*1160*	*889*	*278*
4402	*2419*	*301*	*-200*	*-20*	*1160*	*867*	*277*
4383	*2420*	*279*	*-196*	*-12*	*1162*	*873*	*278*
8690	*2420*	*286*	*-208*	*9*	*1161*	*866*	*273*
1483	*2421*	*277*	*-184*	*-25*	*1162*	*880*	*280*
1mam							
4204	**3173**	**332**	**-217**	**-14**	**1303**	**1452**	**287**
4093	**3455**	**645**	**-279**	**-9**	**1305**	**1450**	**313**
4203	**3183**	**326**	**-204**	**-14**	**1303**	**1453**	**289**
4218	**3181**	**326**	**-208**	**-7**	**1302**	**1449**	**288**
4217	**3181**	**325**	**-207**	**-8**	**1302**	**1450**	**288**
868	**3263**	**422**	**-249**	**16**	**1307**	**1452**	**283**
3238	**3140**	**393**	**-259**	**-21**	**1304**	**1408**	**283**
3317	*3099*	*301*	*-220*	*7*	*1303*	*1402*	*273*
88	*3115*	*305*	*-210*	*11*	*1304*	*1399*	*273*
87	*3115*	*304*	*-208*	*8*	*1304*	*1400*	*275*
3319	*3105*	*307*	*-219*	*7*	*1304*	*1402*	*273*
89	*3126*	*310*	*-210*	*12*	*1304*	*1404*	*274*
3173	*3114*	*369*	*-250*	*-17*	*1302*	*1399*	*279*
2975	*3130*	*325*	*-220*	*-4*	*1305*	*1409*	*285*

1vfa

4781	**2944**	**1205**	**-450**	**-83**	**1150**	**922**	**185**
4782	**3012**	**1291**	**-474**	**-82**	**1152**	**924**	**186**
4795	**2847**	**1108**	**-420**	**-103**	**1146**	**920**	**180**
1717	**2359**	**642**	**-382**	**-128**	**1138**	**898**	**176**
1715	**2351**	**657**	**-382**	**-128**	**1117**	**901**	**170**
4431	**2178**	**459**	**-305**	**-134**	**1115**	**867**	**161**
1964	**2310**	**599**	**-332**	**-145**	**1129**	**884**	**161**
1779	*2079*	*369*	*-265*	*-154*	*1114*	*849*	*153*
4599	*2089*	*392*	*-295*	*-141*	*1115*	*851*	*153*
395	*2099*	*382*	*-270*	*-155*	*1118*	*852*	*158*
3176	*2107*	*431*	*-299*	*-189*	*1116*	*867*	*168*
4569	*2088*	*381*	*-282*	*-146*	*1116*	*854*	*153*
4461	*2097*	*372*	*-275*	*-155*	*1118*	*865*	*159*
4582	*2096*	*403*	*-299*	*-140*	*1113*	*852*	*153*

1igf

240	**2819**	**645**	**-397**	**30**	**1116**	**1132**	**277**
239	**2813**	**675**	**-418**	**10**	**1112**	**1131**	**276**
242	**2816**	**674**	**-412**	**17**	**1112**	**1125**	**275**
245	**2844**	**716**	**-423**	**16**	**1111**	**1120**	**277**
246	**2837**	**693**	**-410**	**6**	**1111**	**1127**	**283**
1862	**2898**	**727**	**-425**	**36**	**1116**	**1144**	**275**
224	**3558**	**1424**	**-542**	**25**	**1160**	**1166**	**296**
2037	*2698*	*553*	*-374*	*10*	*1114*	*1108*	*261*
2028	*2692*	*541*	*-369*	*11*	*1113*	*1109*	*261*
1989	*2700*	*538*	*-363*	*0*	*1117*	*1120*	*264*
28	*2721*	*566*	*-381*	*9*	*1113*	*1122*	*266*
186	*2726*	*564*	*-376*	*8*	*1115*	*1126*	*264*
306	*2738*	*639*	*-430*	*6*	*1112*	*1119*	*268*
1999	*2709*	*538*	*-366*	*15*	*1116*	*1118*	*264*

as well as the framework) interaction energy:

- H3 backbone/FR backbone
- H3 backbone/FR sidechain
- H3 sidechain/FR backbone
- H3 sidechain/FR sidechain

The intra-H3 energy was partitioned into backbone/backbone, backbone/sidechain and sidechain/sidechain.

Table 11 shows that in general, for the H3/FR interactions, those involving H3 sidechains, and particularly sidechain/sidechain interactions, contributed most to the high energy of the low RMSD set. The differences in energy compared to the low energy set were also the greatest here; the sidechain/sidechain interactions were 20kcal higher in the low RMSD/high energy than the low energy conformation set, and in some cases were over 200kcal more. The intra-H3 sidechain interactions were also higher in the high energy set.

The H3 sidechain interactions were not the whole story, however, as the H3/H3 backbone interactions differed between the low and high energy conformation sets by 20-50kcal. The next sub-section investigates this.

Table 11. Breakdown of the repulsive van der Waals energy into intra-H3 and H3-framework/CDR components, and into backbone-backbone, backbone-sidechain and sidechain-sidechain components. Energies are in kcal/mole, and represent the sum of all the interactions involving H3 greater than 1kcal/mole (as we are interested in locating the high energy interactions). For the H3-FR/CDR interactions, the FR-CDR component comes first in the abbreviations, such that bb-sc indicates framework backbone with H3 sidechain, and sc-bb vice-versa. High energy conformations are in bold; low energy conformations in italic.

Conf	H3-H3	H3-FR/ other CDR	bb-bb	H3-FR/ bb-sc	other sc-bb	CDR sc-sc	bb-bb	H3-H3 bb-sc	sc-sc
1cgs									
7145	**255**	**29**	**0**	**18**	**0**	**11**	**57**	**111**	**87**
4071	**164**	**13**	**0**	**0**	**0**	**13**	**91**	**58**	**16**
4072	**165**	**13**	**0**	**0**	**0**	**13**	**93**	**56**	**16**
4070	**162**	**17**	**0**	**0**	**1**	**16**	**86**	**62**	**14**
1318	**161**	**16**	**0**	**0**	**11**	**5**	**96**	**54**	**10**
5284	**161**	**34**	**0**	**0**	**7**	**27**	**73**	**60**	**28**
1397	**225**	**192**	**0**	**36**	**31**	**124**	**82**	**97**	**47**
4129	*128*	*5*	*0*	*0*	*1*	*3*	*57*	*52*	*19*
4375	*119*	*11*	*0*	*0*	*7*	*4*	*61*	*46*	*12*
2952	*143*	*11*	*0*	*0*	*0*	*11*	*46*	*73*	*23*
4402	*139*	*12*	*0*	*0*	*3*	*9*	*55*	*69*	*16*
4383	*122*	*9*	*0*	*0*	*7*	*2*	*64*	*50*	*8*
8690	*125*	*10*	*0*	*4*	*2*	*4*	*62*	*51*	*12*
1483	*121*	*9*	*0*	*0*	*0*	*0*	*62*	*48*	*11*
1vfa									
4781	**540**	**281**	**2**	**30**	**0**	**249**	**135**	**193**	**212**
4782	**574**	**304**	**2**	**47**	**0**	**254**	**121**	**253**	**200**
4795	**529**	**230**	**2**	**9**	**0**	**219**	**120**	**194**	**215**
1717	**277**	**61**	**2**	**37**	**0**	**22**	**77**	**154**	**46**
1715	**290**	**66**	**2**	**36**	**5**	**23**	**68**	**153**	**69**
4431	**193**	**34**	**1**	**2**	**0**	**30**	**57**	**90**	**46**
1964	**274**	**64**	**1**	**4**	**0**	**59**	**78**	**109**	**88**
1779	*169*	*6*	*6*	*0*	*0*	*0*	*74*	*58*	*38*
4599	*167*	*11*	*2*	*0*	*0*	*9*	*73*	*63*	*31*
395	*159*	*11*	*2*	*0*	*0*	*9*	*63*	*61*	*37*
3176	*191*	*11*	*1*	*2*	*0*	*8*	*74*	*69*	*48*
4569	*163*	*7*	*2*	*0*	*0*	*5*	*67*	*71*	*26*
4461	*170*	*5*	*2*	*0*	*0*	*4*	*61*	*64*	*45*
4582	*167*	*22*	*2*	*10*	*0*	*10*	*63*	*55*	*49*

Conf	H3-H3	H3-FR/ other CDR	bb-bb	H3-FR/ bb-sc	other sc-bb	CDR sc-sc	bb-bb	H3-H3 bb-sc	sc-sc
1igf									
240	**329**	**31**	**0**	**7**	**0**	**24**	**143**	**126**	**60**
239	**324**	**31**	**1**	**1**	**0**	**28**	**134**	**128**	**63**
242	**329**	**32**	**0**	**12**	**0**	**20**	**142**	**142**	**45**
245	**358**	**38**	**0**	**0**	**0**	**38**	**124**	**147**	**88**
246	**338**	**45**	**0**	**0**	**0**	**45**	**115**	**144**	**79**
1862	**327**	**76**	**1**	**48**	**5**	**22**	**150**	**129**	**48**
224	**466**	**485**	**2**	**66**	**0**	**417**	**142**	**189**	**35**
2037	*243*	*28*	*0*	*12*	*2*	*13*	*107*	*84*	*52*
2028	*241*	*25*	*0*	*10*	*2*	*12*	*101*	*82*	*58*
1989	*253*	*6*	*0*	*0*	*3*	*3*	*85*	*97*	*71*
28	*255*	*19*	*0*	*0*	*4*	*15*	*75*	*130*	*51*
186	*259*	*22*	*0*	*1*	*5*	*15*	*83*	*121*	*45*
306	*265*	*47*	*2*	*24*	*9*	*12*	*85*	*111*	*69*
1999	*240*	*19*	*0*	*3*	*3*	*13*	*102*	*94*	*43*

Intra-H3 energies

The H3/H3 backbone interactions above 5kcal were investigated; they are listed in Table 12. In addition, for 1mam and 1vfa, angle bending interactions above 1kcal were investigated (Table 12a), as angle bending energies differed considerably between low and high overall energy conformations for these structures.

The results show that 1,4-interactions (Figure 16) constitute the majority of repulsive intra-H3 backbone/backbone interactions above 5kcal. These are concentrated at the join of the loop to the framework (as they were in unminimised structures - see Chapter 2), as are the high angle energies for the low RMSD/high energy conformations of

Table 12. Backbone-backbone intra-H3 interactions (including interactions with the residues either side of the loop) over 5kcal/mole for the high and low energy minimised conformations analysed in chapter 4. 1,4 interactions are shown in bold; 1,4 interactions at the loop/framework join are shown in bold and italics.

1cgs H3=213-219

High energy Conf	Interaction	Energy	Low energy Conf	Interaction	Energy
7145	**217O/218CA** ***219N/220N***	**5.03** ***8.27***	7129	-	-
4071	-	-	7124	-	-
4072	**215N/216N**	**5.00**	7128	-	-
4070	-	-	7143	-	-
1318	212O/218N ***219CB/220N***	7.60 ***10.52***	7120	-	-
5284	**215N/216N** ***219CB/220N***	**6.13** ***8.37***	7655	-	-
1397	212O/218CA **217CB/218N** **218O/219CA**	8.53 **17.87** **9.18**	3323	212O/218CA	5.16

1vfa H3=206-213

High energy Conf	Interaction	Energy	Low energy Conf	Interaction	Energy
4781	**208C/209C**	**9.03**	1755	**212N/213N**	**5.63**
	209C/210CB	**5.14**			
	210O/211CA	**5.01**			
	213N/214N	**5.38**			
	213CB/213O	**10.03**			
	213O/214CA	***9.34***			
4782	**208C/209C**	**7.71**	1779	**212N/213N**	**5.73**
	209C/210CB	**5.12**			
	213N/214N	**5.43**			
	213CB/213O	**10.26**			
	213O/214CA	***9.33***			
4795	**207CB/208N**	**5.17**	2491	-	-
	208C/209C	**5.54**			
	208O/209C	**6.07**			
	213N/214N	***5.36***			
	213CB/213O	***9.94***			
	213O/214CA	***9.34***			
1717	**205C/206CB**	**5.28**	2482	**209CB/210N**	**6.18**
	207CB/208N	**6.54**			
	208C/209CB	**5.20**			
	208O/209CB	5.95			
	211N/212N	**6.40**			
	211O/214NE1	5.98			
	212CA/214CD1	9.49			
	212C/214CD1	5.33			
1715	211O/214NE1	11.07	2454	-	-
	205C/206CB	**5.88**			
	207CB/208N	**6.49**			
	208C/209CB	**5.32**			
	208O/209CB	6.16			
	210O/211CA	**5.63**			
	212CA/214CD1	9.05			
	212C/214CD1	6.05			
4431	**207CB/208N**	**5.01**	4582	**209CB/210N**	**5.26**
1964	**212N/213N**	**5.82**	395	-	-
	213O/214CA	***5.86***			

1igf H3=213-222

High energy Conf	Interaction	Energy	Low energy Conf	Interaction	Energy
240	**218N/219N**	**7.33**	2037	-	-
	222CB/223N	***10.74***			
239	***222CB/223N***	***10.89***	2028	-	-
242	***222CB/223N***	***10.79***	1899	-	-
245	212O/221N	5.43	1770	-	-
	218CB/219N	**5.66**			
	222CB/223N	***10.63***			
246	***222CB/223N***	***11.21***	1887	**218CB/219N**	**6.98**
1862	212O/221N	8.30	309	-	-
	212O/221O	5.47			
	213N/214N	**5.97**			
	218N/219N	**8.38**			
	221O/222CA	**5.36**			
	222CB/223N	***14.57***			
224	212O/221N	6.89	27	-	-
	212O/221O	7.04			
	218O/219CB	**5.02**			
	219CB/220N	**6.63**			
	222CB/223N	***13.39***			

1mam H3=210-217

High energy Conf	Interaction	Energy	Low energy Conf	Interaction	Energy
4204	-	-	3317	-	-
4093	210CB/215O	6.88	88	-	-
	216C/217CB	**6.66**			
	217CB/217O	**5.19**			
	217O/218CA	***5.28***			
4203	***217CB/218N***	***8.59***	87	-	-
4218	***217CB/218N***	***8.56***	3319	-	-
4217	***217CB/218N***	***8.51***	89	-	-
868	209NH1/215CB	5.92	**3173**	**211N/212N**	**8.27**
	209C/210CB	***8.30***			
	210CB/211N	**15.60**			
	211N/212N	**13.16**			
3238	209NH1/215O	6.45	**2975**	**213CA/214CA**	**5.19**
	211N/212N	**8.29**			

Table 12a. Bond angles above 5kcal/mole in the minimised structures (1mam and 1vfa only). These high energy angles were found only in the high energy set of structures, so no low energy structures are shown. Angles at the loop/framework join are shown in bold.

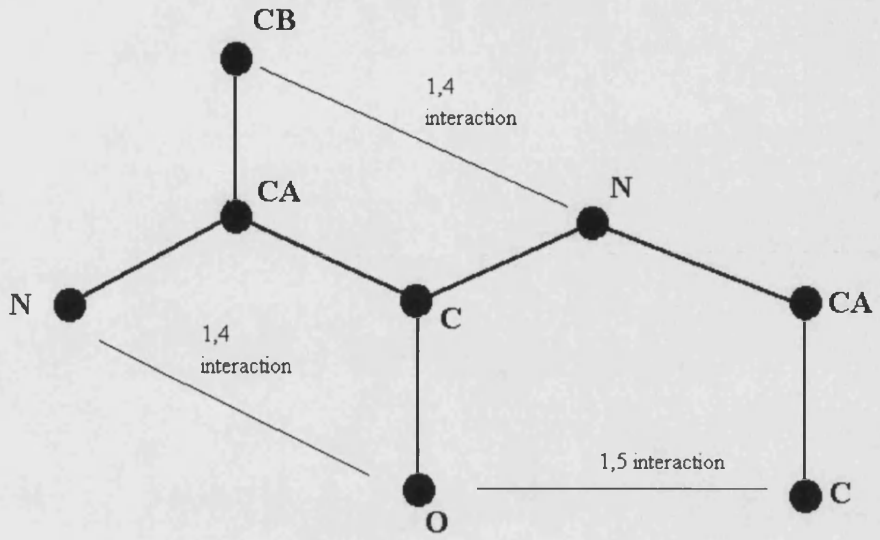
1mam H3=210-217

Conformation	Angle	Energy
4204	**res 217, N-CA-C**	**6.4**
	res 217, N-CA-CB	**24.1**
4218	**res 217, N-CA-C**	**5.9**
	res 217, N-CA-CB	**23.8**
868	**res 210, N-CA-C**	**19.0**
	res 210, N-CA-CB	**8.0**

1vfa H3=206-213

1717	**res 206, N-CA-C**	**5.0**
	res 206, N-CA-CB	**18.2**
1715	**res 206, N-CA-C**	**5.3**
	res 206, N-CA-CB	**18.9**
4431	**res 213, CB-CA-C**	**6.9**

1mam and 1vfa, although there are 1,4-interactions in other regions of the loops.



1,4 and 1,5 interactions

Figure 16. 1,4 and 1,5 interactions in a section of peptide.

High energy interactions: artefacts of minimisation?

One further investigation that is of value is to examine the location of the high energy intra-H3 interactions in the unminimised conformations. This is because minimisation can distribute a local high energy interaction throughout the loop, which could mean that the 1,4 interactions are merely artefacts of minimisation. It could also be disguising a localised site of 1,4 interactions by distributing it

throughout the loop. Table 13 shows the repulsive van der Waals interactions above 30kcal for each structure.

Table 13. The van der Waals atom/atom interactions above 30kcal/mole for the non-minimised conformations. 1,4 interactions are shown in bold; note that all interactions betweem adjacent carbonyls are considered 1,4 interactions as the carbonyl is rigid - it is treated here in inter-carbonyl interactions as a single entity.The 'Type' of interaction refers to the four principal sites of high energy: 1=at the loop/framework join; 2=in the chain closure region; 3=at the database/CONGEN join; 4=between the backbones at each end of the loop.
In interactions involving several atoms of a proline ring and several adjacent backbone atoms or a carbonyl group, the interactions are grouped together using the abbreviations PR=proline,BB=backbone atoms and CO=carbonyl.

1cgs H3=213-219

High	energy	set		Low	energy	set	
Conf	*7145*			*Conf*	*4129*		
Atom1	Atom2	Energy	Type	Atom1	Atom2	Energy	Type
216O	**217CB**	**78**	**-**	**212O**	**218N**	**31**	**4**
215N	**216N**	**48**	**2**	**214N**	**215N**	**46**	**3**
216O	**217CB**	**78**	**2**	**215N**	**216N**	**39**	**2**
217N	**218N**	**34**	**2**	**218N**	**219N**	**32**	**1**
219N	**220N**	**58**	**1**	**219N**	**219O**	**30**	**1**
Conf	*4071*			*Conf*	*4375*		
212O	218N	37	4	**214N**	**215N**	**41**	**3**
212O	218O	549	4	**215N**	**216N**	**54**	**2**
213O	217N	102	-	**217N**	**218N**	**31**	**1**
215N	**216N**	**55**	**2**				
216O	218N	40	-				
218C	**219C**	**36**	**1**				
219CB	220N	40966	1				
Conf	*4072*			*Conf*	*2952*		
212O	218N	37	4	**214N**	**215N**	**46**	**3**
212O	218O	549	4	**219CB**	**220N**	**537**	**1**
215N	**216N**	**35**	**2**				
216O	218N	40	-				
218C	**219C**	**36**	**1**				
219CB	220N	40966	1				

1cgs H3=213-219

High	energy	set		Low	energy	set	
Conf	*4070*			*Conf*	*4402*		
Atom1	Atom2	Energy	Type	Atom1	Atom2	Energy	Type
212O	218N	37	4	**214N**	**215N**	**39**	**3**
212O	218O	549	4	**218N**	**219N**	**41**	**1**
215N	**216N**	**35**	**2**				
216N	**217N**	**33**	**2**				
216C	**217C**	**34**	**2**				
216O	**217O**	**59**	**2**				
218C	**219C**	**36**	**1**				
219CB	**220N**	**40966**	**1**				
Conf	*1318*			*Conf*	*4383*		
212N	218O	33	4	**214N**	**215N**	**46**	**3**
212O	218O	661	4	**215N**	**216N**	**46**	**2**
213N	**213O**	**30**	**-**	216O	217CB	30	-
214N	**215N**	**46**	**3**	**217N**	**218N**	**37**	**1**
215N	**216N**	**43**	**2**				
216N	**217N**	**38**	**2**				
219CB	**220N**	**397**	**1**				
Conf	*5284*			*Conf*	*8690*		
212O	218O	42	4	**214N**	**215N**	**39**	**3**
215N	**216N**	**50**	**2**	**215N**	**216N**	**43**	**2**
218C	**219C**	**33**	**1**				
219CB	**220N**	**1397**	**1**				
Conf	*1397*			*Conf*	*1483*		
214N	**215N**	**31**	**3**	212O	218O	31	4
216O	**217O**	**33**	**2**	**214N**	**215N**	**39**	**3**
219N	**219O**	**31**	**1**	**216O**	**217O**	**33**	**2**
				219CB	**220N**	**478**	**1**

1mam H3=210-217

High	energy	set		Low	energy	set	
Conf	*4204*			*Conf*	*3317*		
Atom1	Atom2	Energy	Type	Atom1	Atom2	Energy	Type
210CB	215CB	53	-				
212N	**213N**	**37**	**3**				
212O	214PR	160	-				
212CB	**218N**	**30187**	**1**				
Conf	*4093*			*Conf*	*88*		
209C	**210CB**	**15984**	**1**	**210N**	**210O**	**33**	**1**
209O	210CB	14915	1	**212N**	**213N**	**39**	**3**
210BB	211PR	3500	-				
210O	212N	30	-				
212N	**213N**	**39**	**3**				
213N	214PR	42	-				
217CB	**217O**	**10^6**	**1**				
Conf	*4203*			*Conf*	*87*		
212N	**213N**	**37**	**3**	**21ON**	**210O**	**33**	**1**
212CO	214PR	132	-	**212N**	**213N**	**44**	**3**
213CO	214PR	155	-				
213O	215O	62	-				
217CB	**218N**	**30187**	**1**				
Conf	*4218*			*Conf*	*3319*		
214C	**215C**	**43**	**2**	212O	215N	39	-
214O	**215C**	**54**	**2**				
214O	**215O**	**131**	**2**				
217CB	**218N**	**30187**	**1**				
Conf	*4217*			*Conf*	*89*		
213N	214PR	72	-	**210N**	**210O**	**33**	**1**
214C	**215C**	**34**	**2**	**212N**	**213N**	**31**	**3**
214O	**215O**	**43**	**2**	212O	214PR	118	-
217CB	**218N**	**30187**	**1**	**214C**	**215C**	**34**	**3**
				214O	**215O**	**147**	**3**
Conf	*868*			*Conf*	*3173*		
209C	**210CB**	**3482**	**1**	**209C**	**210CB**	**7821**	**1**
213C	**214C**	**34**	**2**	**209C**	**210C**	**32**	**1**
213O	**214O**	**111**	**2**	**217CB**	**217O**	**60**	**1**
216N	**217N**	**47**	**3**				
Conf	*3238*			*Conf*	*2975*		
209C	**210CB**	**7821**	**1**	**209C**	**210CB**	**51**	**1**
209C	**210C**	**32**	**1**	209O	216O	115	4
215N	**216N**	**44**	**2**	**212N**	**213N**	**44**	**3**
217CB	**217O**	**60**	**1**	**216N**	**217N**	**32**	**3**
				217CB	**218N**	**10^6**	**1**
				217CB	218CA	31	1

1vfa H3=206-213

High	energy	set		Low	energy	set	
Conf	*4781*			*Conf*	*1779*		
Atom1	Atom2	Energy	Type	Atom1	Atom2	Energy	Type
C207	CA216	166	4	N206	CA207	53	1
CB208	**O208**	**80**	**-**	C207	C210	53	-
CB208	CA209	60	-	**CB208**	**O208**	**50**	**-**
C208	**C209**	**144**	**2**	CB208	CA209	60	-
N209	**O209**	**185**	**2**	**C208**	**C209**	**144**	**2**
N209	**N210**	**500**	**2**	**N209**	**O209**	**185**	**2**
CB209	**N210**	**84**	**2**	**N209**	**N210**	**499**	**2**
CB210	**O210**	**8079**	**2**	**CB209**	**N210**	**84**	**2**
O210	**C211**	**5799**	**2**	**CB210**	**O210**	**8067**	**2**
N212	N214	77	1	**O210**	**C211**	**5797**	**2**
CA212	N214	1366	1	**O211**	**C212**	**52**	**3**
CB213	**O213**	**257**	**1**	N212	N214	75	1
				CA212	N214	13119	1
				CB213	**O213**	**257**	**1**
Conf	*4782*			*Conf*	*4599*		
C207	CA216	225	4	C207	C210	229	-
O207	CA216	9858	4	**CB208**	**O208**	**80**	**-**
O210	**C211**	**5799**	**2**	CB208	CA209	60	-
CB208	**O208**	**80**	**-**	**C208**	**C209**	**144**	**2**
CB208	CA209	60	-	**N209**	**O209**	**185**	**2**
C208	**C209**	**144**	**2**	**N209**	**N210**	**500**	**2**
N209	**O209**	**185**	**2**	**CB209**	**N210**	**84**	**2**
N209	**N210**	**500**	**2**	**CB210**	**O210**	**8059**	**2**
CB209	**N210**	**84**	**2**	**O210**	**C211**	**5801**	**2**
CB210	**O210**	**8079**	**2**	N212	N214	75	1
N212	N214	77	1	CA212	N214	13115	1
CA212	N214	13068	1	**CB213**	**O213**	**258**	**1**
CB213	**O213**	**257**	**1**				

1vfa H3=206-213

High	energy	set		Low	energy	set	
Conf	*4795*			*Conf*	*395*		
Atom1	Atom2	Energy	Type	Atom1	Atom2	Energy	Type
C207	CA210	188	-	C207	CA210	50	-
CB208	**O208**	**80**	**-**	**CB208**	**O208**	**80**	**-**
CB208	CA209	60	-	**CB208**	**CA209**	**60**	**2**
C208	**C209**	**144**	**2**	**C208**	**C209**	**145**	**2**
N209	**O209**	**185**	**2**	**N209**	**O209**	**185**	**2**
N209	**N210**	**500**	**2**	**N209**	**N210**	**501**	**2**
CB209	**N210**	**84**	**2**	**CB209**	**N210**	**84**	**2**
CB210	**O210**	**8079**	**2**	**CB209**	**O210**	**8043**	**2**
O210	**C211**	**8800**	**2**	**O210**	**C211**	**5825**	**2**
O211	N214	99	1	N212	N214	96	1
N212	N214	175	1	CA212	N214	13076	1
CA212	N214	3151	1	**CB213**	**O213**	**257**	**1**
CB213	**O213**	**256**	**1**				
Conf	*1717*			*Conf*	*3176*		
C205	**CB206**	**541**	**1**	**C205**	**CB206**	**145**	**1**
O206	**O207**	**1084**	**-**	C207	C210	351	-
N207	**O207**	**91**	**-**	O207	C210	77	-
CB208	**O208**	**80**	**-**	**CB208**	**O208**	**80**	**-**
CB208	CA209	60	-	CB208	CA209	60	-
C208	**C209**	**145**	**2**	**C208**	**C209**	**145**	**2**
N209	**O209**	**135**	**2**	**N209**	**O209**	**185**	**2**
CB209	**N210**	**84**	**2**	**N209**	**N210**	**500**	**2**
CB210	**O210**	**8028**	**2**	**CB209**	**N210**	**84**	**2**
O210	**C211**	**5813**	**2**	**CB210**	**O210**	**8037**	**2**
C210	O214	107	1	**O210**	**C211**	**5818**	**2**
N212	N214	67	1	N212	N214	75	1
CA212	N214	13145	1	CA212	N214	13193	1
CB213	**O213**	**258**	**1**	**CB213**	**O213**	**257**	**1**

1vfa H3=206-213

High	energy	set		Low	energy	set	
Conf	*1715*			*Conf*	*4569*		
Atom1	Atom2	Energy	Type	Atom1	Atom2	Energy	Type
C205	**CB206**	**541**	**1**	C207	C210	229	-
O206	**O207**	**1084**	**-**	**CB208**	**O208**	**80**	**-**
N207	**O207**	**91**	**-**	CB208	CA209	60	-
CB208	**O208**	**80**	**-**	**C208**	**C209**	**144**	**2**
CB209	CA209	60	-	**N209**	**O209**	**185**	**2**
C208	**C209**	**145**	**2**	**N209**	**N210**	**500**	**2**
N209	**O209**	**185**	**2**	**CB209**	**N210**	**84**	**2**
N209	**N210**	**500**	**2**	**CB210**	**O210**	**859**	**2**
CB209	**N210**	**84**	**2**	**O210**	**C211**	**5801**	**2**
CB210	**O210**	**808**	**2**	N212	N214	74	1
O210	**C211**	**5813**	**2**	CA212	N214	13157	1
N212	N214	64	1	**CB213**	**O213**	**257**	**1**
CA212	N214	13223	1				
CB213	**O213**	**256**	**1**				

Conf	*4431*			*Conf*	*4461*		
CB208	**O208**	**79**	**-**	N207	N215	144	4
CB208	CA209	60	-	**CB208**	**O208**	**79**	**-**
C208	**C209**	**144**	**2**	CB208	CA209	60	-
N209	**O209**	**185**	**2**	**C208**	**C209**	**144**	**2**
N209	**N210**	**498**	**2**	**N209**	**O209**	**185**	**2**
CB209	**N210**	**84**	**2**	**N209**	**N210**	**498**	**2**
CB210	**O210**	**8065**	**2**	**CB209**	**N210**	**84**	**2**
O210	**C211**	**5831**	**2**	**CB210**	**O210**	**8065**	**2**
O211	N213	57	-	**O210**	**C211**	**5811**	**2**
N212	N214	75	1	O211	N214	78	1
CA212	N214	13146	1	O211	CA214	123	1
CB213	**O213**	**256**	**1**	N212	N214	75	1
				CA212	N214	13124	1
				CB213	**O213**	**256**	**1**

1vfa H3=206-213

High	energy	set		Low	energy	set	
Conf	*1964*			*Conf*	*4582*		
Atom1	Atom2	Energy	Type	Atom1	Atom2	Energy	Type
C206	C210	126	-	C207	C210	229	-
C207	C210	122	-	**CB208**	**O208**	**80**	**-**
CB208	CA209	61	-	CB208	CA209	60	-
C208	**C209**	**144**	**2**	**C208**	**C209**	**144**	**2**
N209	**O209**	**185**	**2**	**N209**	**O209**	**185**	**2**
N209	**N210**	**84**	**2**	**N209**	**N210**	**84**	**2**
CB210	**O210**	**8039**	**2**	**CB209**	**N210**	**84**	**2**
O210	**C211**	**5808**	**2**	**CB210**	**O210**	**8059**	**2**
O211	N214	73	1	**O210**	**C211**	**5801**	**2**
N212	N214	75	1	N212	N214	75	1
CA212	N214	13242	1	CA212	N214	13166	1
CB213	**O213**	**257**	**1**	**CB213**	**O213**	**257**	**1**

1igf H3=213-222

High	energy	set		Low	energy	set	
Conf	*240*			*Conf*	*2037*		
Atom1	Atom2	Energy	Type	Atom1	Atom2	Energy	Type
212O	221N	48	4	**215N**	**216N**	**35**	**-**
215N	**216N**	**33**	**-**	**218N**	**219N**	**32**	**2**
216O	218N	44	2				
218N	**219N**	**44**	**2**				
220N	**220O**	**32**	**3**				
221N	**221O**	**40**	**-**				
221O	**222O**	**30**	**1**				
222CB	**223N**	**175**	**1**				
Conf	*239*			*Conf*	*2028*		
212O	221N	48	4	**215N**	**216N**	**35**	**-**
214CB	219CB	33	-	216CB	217CD	34	-
215N	**216N**	**33**	**-**	**216O**	**217O**	**46**	**2**
216C	219N	52	-				
217C	219N	77	2				
218N	**219N**	**39**	**2**				
221N	**2210**	**40**	**-**				
221O	**222O**	**30**	**1**				
222CB	**223N**	**175**	**1**				
Conf	*242*			*Conf*	*1989*		
212O	221N	48	4	**215N**	**216N**	**35**	**-**
214N	219O	73	-	**216C**	**217C**	**31**	**2**
215N	**216N**	**33**	**-**	**217O**	**218CB**	**45**	**2**
216O	218N	122	2				
218N	219N	33	2				
220N	**220O**	**32**	**3**				
221N	**221O**	**40**	**-**				
221O	**222O**	**30**	**1**				
222CB	**223N**	**175**	**1**				
Conf	*245*			*Conf*	*28*		
212O	221N	48	4	**215N**	**216N**	**40**	**-**
215N	**216N**	**33**	**-**	**218O**	**219CB**	**47**	**2**
22ON	**22O0**	**38**	**3**	**221N**	**222N**	**32**	**-**
221N	**221O**	**40**	**-**				
221O	**222O**	**30**	**1**				
222CB	**223N**	**175**	**1**				
Conf	*246*			*Conf*	*186*		
212O	221N	48	4	**215N**	**216N**	**30**	**-**
215N	**216N**	**33**	**-**				
220N	**220O**	**32**	**3**				
221N	**221O**	**40**	**-**				
221O	**222O**	**30**	**1**				
222CB	**223N**	**175**	**1**				

1igf H3=213-222

High	energy	set		Low	energy	set	
Conf	*1862*			*Conf*	*306*		
Atom1	Atom2	Energy	Type	Atom1	Atom2	Energy	Type
212O	221N	51	4	**217O**	**218CB**	**43**	**2**
212O	221O	89	4	**218N**	**219N**	**35**	**2**
213N	**214N**	**44**	**-**	**221N**	**222N**	**35**	**-**
214N	**215N**	**42**	**-**				
215N	**216N**	**32**	**-**				
216CB	217PR	52	-				
217CD	218N	36	-				
218N	**219N**	**42**	**2**				
221N	**221O**	**30**	**-**				
222CB	**223N**	**177**	**1**				
Conf	*224*			*Conf*	*1999*		
212O	221N	48	4	**215N**	**216N**	**35**	**-**
215N	**216N**	**33**	**-**	**217O**	**218CB**	**35**	**2**
216O	**217O**	**117**	**2**	**218N**	**219N**	**30**	**2**
217C	**218C**	**31**	**2**				
218O	**219CB**	**52**	**2**				
220N	**220O**	**38**	**3**				
221N	**221O**	**40**	**-**				
221O	**222O**	**30**	**1**				
222CB	**223N**	**175**	**2**				

The Table confirms that 1,4 interactions are a genuine problem, as they are present in the unminimised high energy/low rms conformations. As in the minimised structures, they are also concentrated around the loop/framework join, and particularly between the C-beta of the loop C-terminal residue and both the carbonyl oxygen of the loop C-terminal residue and the backbone nitrogen atom of the residue following - recall that the C-alpha and carbonyl of the C-terminal loop residue are treated by CAMAL as framework atoms. These 'join' interactions have the highest energy, in some cases above 1000kcal. Indeed, the high energy at the joins mentioned in Chapter 2 arises largely from this 1,4 interaction. A secondary site of high energy 1,4 interactions can also be seen, in the centre of the loop, corresponding to the region built using the Go and Scheraga chain closure algorithm. There are also some high energy 1,5 interactions in these two sites, though these are lower energy than the 1,4-interactions. A third, less important, site of high energy 1,4 interactions is at the join between the database and CONGEN constructed region of the loops.

Finally, there is a non-1,4 high energy interaction that is frequently seen, between backbone atoms at each end of the loop.

Conclusions

Breaking down the high energy interactions observed in certain low RMSD conformations has shown that the dominant contibution is repulsive van der Waals, with angle bending energy also contributing in some structures. The repulsive van der Waals has two main high energy components:

i) 1,4 clashes:

These are found throughout the loop but the highest are at the loop/framework join, and to a lesser extent in the chain-closure algorithm region. In particular, the interactions between the C-beta of the C-terminal loop atom (built by CAMAL) and both the carbonyl oxygen of the C- terminal loop residue and the backbone nitrogen atom of the following residue (treated as framework) are frequently high in energy. This suggests a problem with the loop/framework grafting procedure. In addition, a secondary site of high energy 1,4-interactions is found in the Go and Scheraga chain closure region, suggesting a problem with this procedure. There are also a few high energy 1,4-interactions in other regions of the loop.
While the 1,4 interactions remain relatively high in high energy/low RMSD structures after minimisation, they are reduced in energy from above 50 to below 10 kcal, and are less concentrated (though still

fairly concentrated) in the join region, and more spread out throughout the loop. However, these small energies add up so that the conformations have considerably higher (20-50 kcal) intra- H3 backbone/backbone energy than the lowest energy loops.

ii) H3 sidechain clashes:

These contribute most of the energy to the H3/framework interaction energy, and in some conformations add up to over 200 kcal. They also differ the most between the high and low energy conformation sets (up to 200 kcal). This suggests that H3 sidechains are often misplaced in otherwise good structures by the CONGEN sidechain placement algorithm, raising their energy above poor structures that happen to have sidechains which avoid bad clashes.

iii) Other interactions:

In two structures, 1mam and 1vfa, the loop joins contain high energy angles. In non-minimised structures, the grafting procedure leads to clashes between backbone atoms at each end of the loop, though this is reduced on minimisation.

4.3 The effect of removing 1,4 interactions and H3 sidechains

As seen above, the 1,4 interactions and H3 sidechain clashes appear to be the main causes of high energy in low RMSD conformations. The CONGEN and VFF stages of the modelling were altered by deletion of these components to see if any improvement occurred. Removing H3 sidechains has the further advantage that the possibility of 'false' salt-bridges is eliminated.

Method

Eight CDR-H3 loops (the set used in Chapter 2 for the individual terms screen, but not 1bbj, as this was a database loop) were modelled using CAMAL and VFF with 45 rounds of steepest descent

minimisation. The conditions used for each loop were as follows.

Run	H3 sidechains in VFF	1,4s in VFF	Other loop sidechains in CONGEN/VFF
1	Yes	Yes	No
2	No	Yes	Yes
3	Yes	No	No
4	No	No	Yes
5	No	No	No

As seen from the Table above, in runs 1-4, two alterations are made.

i) Removing 1,4-interactions (runs 3 and 4)

ii) Altering the CDR sidechains included in the modelling (runs 2 and 4).

In run 4 not only were H3 sidechains ignored in VFF, but also, the sidechains of the other CDRs were included in CONGEN and VFF to prevent the H3 backbones accessing space that would normally be unavailable to them. If an improvement was seen, it would suggest that their inclusion is important (of course a solution to the problem of modelling Fvs where the sidechain conformations are unknown will eventually have to be found). To assess the effect of including other CDR sidechains AND removing H3 sidechains, versus removing H3 sidechains only, an additional run (5) was carried out for four of the structures, in which the 1,4 interactions and H3 sidechains were removed without including the other CDR sidechains.

In each run, the bottom 200 conformations from VFF were clustered based on RMSD between one conformation and another,

with an initial clustering resolution of 0.5Å, a step size of 0.25Å and a final resolution of 1.0Å (see Appendix 2) and the remaining conformations ranked in terms of energy. The clustering was found to be necessary as many conformations were found to be very similar to each other on visual examination using Insight II.

Note that the sidechains are built in CONGEN whether or not they are included in VFF. This is to act as a filter: without sidechain building in CONGEN, too many conformations (>10000 in most, and >20000 in some cases) are produced. The aim is to rebuild sidechains later, which will be on minimised structures and less likely to lead to clashes (see Chapter 5).

Results

Table 14a and Figure 17 show that in general, using the completely altered set of conditions (run 4) gives improved results, as measured by the rms spread of the conformations within 20kcal of the lowest energy, compared to the original set of conditions (run 1). Three structures, 1mam, 1hil and 1igm display notable improvements; two more, 1igf and 1cgs, improve to some extent and only two, 1for and 1vfa, show a significant deterioration. In 1vfa, this is largely due to conformations outside the bottom 10 (Table 14a), with 6/10 of the bottom 10 with RMSD <2Å, and in 1for, one good (1.3Å RMSD) conformation is in the bottom 10 energies. The data also show that

both removal of 1,4s and alteration of the included CDR sidechains are important; using one alteration without the other (runs 2 and 3) does not show a consistent improvement (even though individual cases do better with only one modification). The spread in run 5 is worse than that for run 4 in two cases, and the same in the other two. This shows that to obtain the improved results, the canonical CDR sidechains need to be included as well as the H3 sidechains removed.

Table 14. The percentage of conformations within 20kcal of the lowest energy conformation in each rms range, for the four different VFF runs followed by clustering ('Old' = H3 sidechains but not other CDR sidechains present, and 'new' vice-versa).

RMSD range	Old + 1,4	New + 1,4	Old - 1,4	New - 1,4
1cgs				
1.0-1.5	0	0	0	0
1.5-2.0	5	0	14	6
2.0-2.5	0	9	4	16
2.5-3.0	16	18	18	22
3.0-3.5	35	36	33	44
3.5-4.0	29	18	20	0
4.0+	15	18	12	13
1mam				
1.0-1.5	0	100	0	56
1.5-2.0	0	0	27	33
2.0-2.5	11	0	23	0
2.5-3.0	11	0	5	0
3.0-3.5	11	0	18	0
3.5-4.0	0	0	0	0
4.0+	67	0	27	11
1vfa				
1.0-1.5	0	0	0	0
1.5-2.0	63	15	75	33
2.0-2.5	25	46	0	24
2.5-3.0	13	0	25	12
3.0-3.5	0	23	0	9
3.5-4.0	0	0	0	9
4.0+	0	15	0	12
2fbj				
1.0-1.5	0	0	4	14
1.5-2.0	64	0	46	43
2.0-2.5	36	100	35	29
2.5-3.0	0	0	8	14
3.0-3.5	0	0	8	0
3.5-4.0	0	0	0	0
4.0+	0	0	0	0
1igf				
1.0-1.5	0	0	0	5
1.5-2.0	25	0	57	45
2.0-2.5	63	33	0	5
2.5-3.0	13	0	29	5
3.0-3.5	0	0	0	15
3.5-4.0	0	0	14	5
4.0+	0	67	0	20

1for

RMSD range	Old + 1,4	New + 1,4	Old - 1,4	New - 1,4
1.0-1.5	7	11	10	6
1.5-2.0	57	56	38	22
2.0-2.5	29	33	41	43
2.5-3.0	7	0	7	16
3.0-3.5	0	0	0	4
3.5-4.0	0	0	0	2
4.0+	0	0	5	6

1hil

1.0-1.5	0	0	0	0
1.5-2.0	0	0	0	4
2.0-2.5	0	17	0	8
2.5-3.0	0	8	16	12
3.0-3.5	0	17	0	31
3.5-4.0	9	8	12	23
4.0+	91	50	72	23

1igm

1.0-1.5	0	0	0	0
1.5-2.0	0	0	0	0
2.0-2.5	0	0	0	43
2.5-3.0	100	0	30	10
3.0-3.5	0	0	26	14
3.5-4.0	0	0	35	29
4.0+	0	100	9	5

Table 14a. The RMSD of the bottom 10 conformations for the 'old' with 1,4s and 'new' without 1,4s runs after clustering ('old' = H3 sidechains but not other CDR sidechains present, and 'new' vice-versa).

1cgs

'Old'		'New'	
Conf	*RMSD*	*Conf*	*RMSD*
2952	4.0	2675	3.0
4383	3.3	2693	2.3
4410	3.5	942	3.0
1483	3.1	44	3.0
2963	4.0	1526	2.5
4370	3.2	4	2.8
8692	3.9	1249	3.1
4395	3.4	2968	4.0
8163	1.8	1387	2.2
8162	1.7	1250	3.0
1mam			
3325	4.3	1538	1.0
3316	4.4	1629	1.6
3196	2.1	1630	1.5
1947	4.8	1737	1.0
87	4.2	1684	1.3
3320	4.4	1570	1.6
88	4.3	1505	1.3
1017	2.7	1577	1.2
2898	3.4	63	4.2
5027	2.4	-	-
1vfa			
1779	1.8	126	1.6
1754	1.6	1155	3.3
1780	1.7	939	1.5
4569	2.2	167	2.4
4452	1.6	119	1.8
4585	2.3	155	2.2
4582	2.5	132	1.5
4461	1.7	1715	1.9
395	2.0	1041	1.5
4457	1.5	1153	3.0

2fbj

'Old'		'New'	
Conf	*RMSD*	*Conf*	*RMSD*
4424	1.9	725	1.8
821	2.4	2075	2.0
6171	1.8	2029	1.0
820	2.4	39	2.5
47	1.9	5	1.7
7040	1.8	1111	2.0
6191	1.7	1938	1.6
819	2.3	795	2.1
4417	2.1	813	2.2
7081	1.9	183	2.3

1igf

2037	1.8	269	1.9
1989	2.6	259	1.9
2040	1.9	29	3.4
2005	2.4	802	2.3
28	2.9	743	1.8
2018	2.0	267	1.7
197	2.4	475	4.4
2029	1.9	63	3.2
189	2.4	90	1.4
2034	2.4	719	3.0

1for

5880	1.4	3141	2.3
675	2.0	3676	2.2
5812	1.7	3137	2.5
5829	1.7	3583	2.1
648	2.6	446	2.5
5760	1.8	218	2.1
306	2.1	3116	2.8
5842	1.9	3440	1.3
676	2.0	3659	2.0
5834	1.6	3147	2.6

1hil

3557	4.3	108	3.4
3590	4.1	5732	3.3
3456	4.2	2074	2.2
3357	4.1	4980	4.8
3366	4.0	5814	3.4
3367	4.0	1660	2.9
3428	4.1	86	3.3
3377	4.3	2017	1.7
3253	4.0	1630	2.6
3586	4.2	362	3.9

1igm

'Old'		'New'	
Conf	*RMSD*	*Conf*	*RMSD*
15814	2.7	3178	3.1
15709	2.9	3344	2.3
15703	2.9	5731	2.0
15737	2.9	2561	3.6
15691	3.1	5338	2.3
15682	3.0	3125	5.4
15670	3.0	5685	2.2
6211	4.3	3144	3.7
7781	3.0	2558	3.6
15798	2.6	5575	2.1

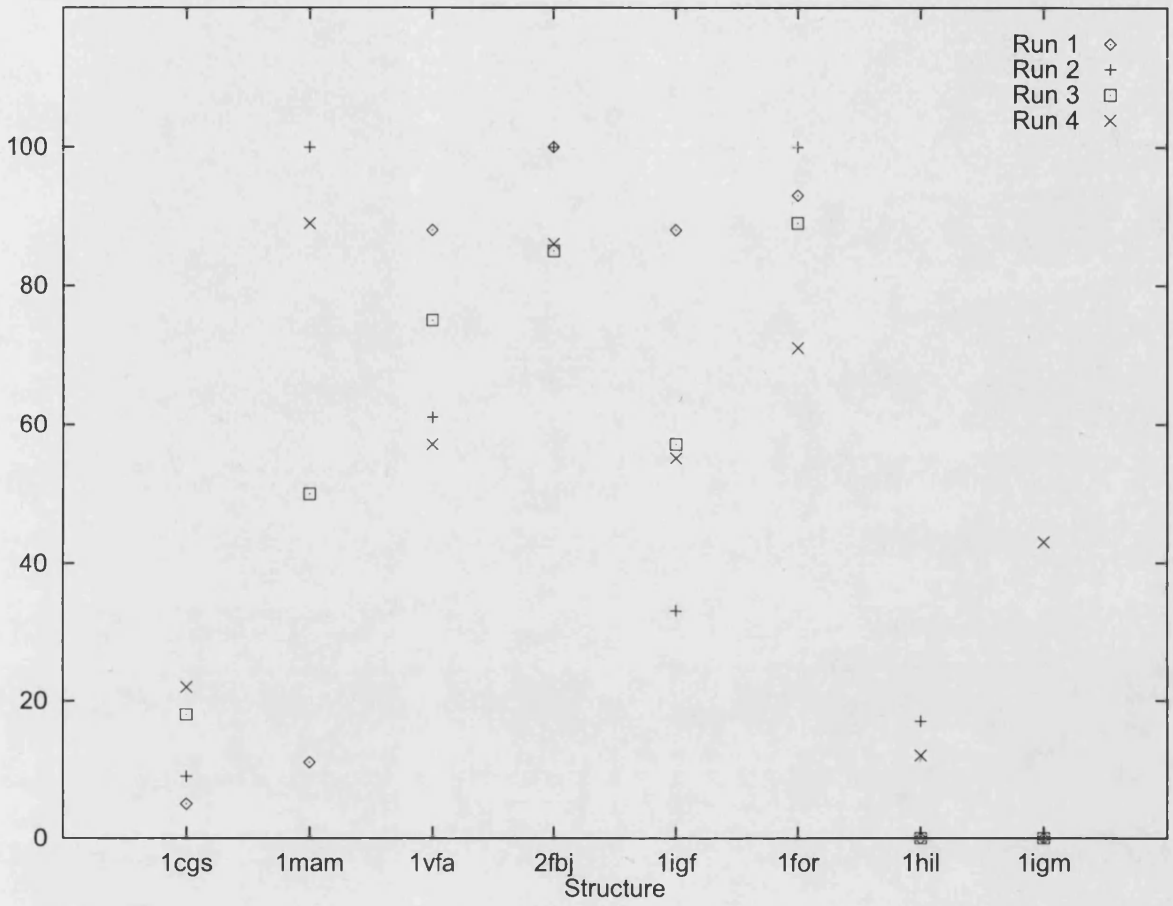


Figure 17. Percentage of H3 conformations within 20kcal/mole of the lowest energy conformation below 2.5 angstroms for the four runs (see text) with and without sidechains and 1,4 interactions.

Using run 4, there is always at least one conformation in the bottom 10 with RMSD no higher than 2.2Å (Table 14b). This has never been seen before; using run 1, two structures (1hil and 1igm) produce only structures with RMSD above 2.5Å. Indeed, using run 1 conditions without clustering, such as was used in Chapter 3 when comparing VFF with solvation screens, three structures (1cgs, 1hil and 1igm) produce only structures with RMSD above 2.5Å.

Another striking improvement is the RMSD spread in the bottom 10 conformations for 1mam (6/10 above 4.0Å in run 1; 8/9 below 2.0Å in run 4). One structure (1for) gives a worse RMSD spread in the bottom 10 in run 4, but as we have seen, there is still a conformation with RMSD 1.3Å. In addition, the lowest RMSD in the bottom 10 in run 4 for 1cgs is 2.2Å, compared to 1.7Å for run 1. On balance, however, since the overall spread within the bottom 10 is no worse, and the spread of the set of conformations within 20kcal of the lowest improves (Table 14a) this is not so important.

Conclusions

It is evident from the results that removing the 1,4 interactions in VFF, removing the CDR-H3 sidechains in VFF and including the canonical CDR sidechains in CONGEN and VFF proved to be a combination of conditions that led to more low energy, low RMSD models, in the majority of cases. However, the low RMSD models

were not always the lowest in energy, leading to a requirement for development of a further screening procedure.

CHAPTER 5 - THE IMPORTANCE OF SIDECHAIN PLACEMENT AND ACCESSIBILITY PATTERNS

5.1 Introduction

The revised modelling procedure, namely removing 1,4 interactions and H3 sidechains from VFF, and including the sidechains of other CDRs in CONGEN and VFF, has been shown to give lower RMSD models in the bottom 10 energy loops than the standard procedure. The lowest energy model of all, however, does not necessarily have low RMSD. This Chapter investigates a number of screening methods to find the most effective means of selecting the low RMSD conformations.

5.2 Screening background

Sidechain placement

Once a relatively small set of low energy backbones is defined, sidechains can be added. This serves two purposes; as well as the obvious one of sidechain modelling, it also serves

as a screen for high RMSD loops in which sidechain clashes will be more likely.

There are a number of basic approaches to sidechain modelling. For conserved framework residues, the typical conformation of known structures is used. For less conserved residues, maximum overlap (Snow and Amzel 1986) can be used, in which the sidechain is modelled so that maximum atomic overlap with the original sidechain of the template framework or loop is achieved. This is followed, if necessary, by small adjustments to chi torsion angles to relieve steric clashes. Rotamer libraries are also used (Ponder and Richards 1987), which are libraries of the most commonly observed rotamers of sidechains. The rotamer in a library which gives the lowest energy in the Fv environment is selected to model the sidechain. This, together with conformational searching, has the problem of combinatorial explosion (having to test every possible chi angle of every sidechain). A number of approaches have been used to deal with this problem.

'COMPOSER' (Blundell et al., 1988) uses a more knowledge based approach to model the sidechains. They are modelled using rules defining the probabilities of sidechain orientation in the equivalent position in homologues, and rules about preferred conformations in secondary structure types. When the sidechain orientation is not expected to be

conserved in a given position among the members of the class of protein, the general rules concerning sidechain conformation in helices and sheets are used.

i) The CONGEN iterative method

The method used by CONGEN to place sidechains is an iterative procedure (Bruccoleri and Karplus 1987). The lowest energy conformation of the first CDR sidechain is built, taking into account the backbone and the sidechains not in this CDR. After this, the succeeding sidechains of the CDR are built in the lowest energy conformations, but in addition taking account of the CDR sidechains built so far. After all CDR sidechains have been built, the first CDR sidechain is rebuilt in a new lowest energy conformation, which will be different to the initial version as the other CDR sidechains would now have been built. These steps are then iterated over all CDR sidechains, rebuilding them, and returning to the start of the CDR after the final CDR sidechain, and so on until the energies of each sidechain converge.

ii) The Dead-end Elimination theorem

The Dead-end Elimination theorem (Desmet et al., 1992; Lasters et al., 1995; see the second reference for more details) is a rotamer-library based sidechain addition procedure, designed to prevent the combinatorial explosion which would result in exploring every combination of rotamers.

In its most simple form, if we wish to try and eliminate a rotamer r of residue i, i(r), we use the following inequality:

$$E_{i(r)} + \Sigma_j \min_s E_{i(r)j(s)} > E_{i(t)} + \Sigma_j \max_s E_{i(t)j(s)} \qquad \textbf{(1)}$$

where j represents all the other residues besides i, $E_{i(r)}$ is the self energy of i(r), and $E_{i(r)j(s)}$ is the interaction energy between the rotamers i(r) and j(s).

What criterion (1) is saying is that if the minimum possible interaction energy between i(r) and all the other rotamers is greater than the maximum possible interaction between i(t) and all the other rotamers, i(r) is dead ending with respect to i(t), and can be eliminated. Essentially, the process is trying to favour i(r), and if its interaction energy is still greater than that of i(t), it is eliminated.

This principle is used in a different way in an extended version (Goldstein, 1994) to produce a more effective way of eliminating rotamers:

$$E_{i(r)} - E_{i(t)} + \Sigma_j \min_s [E_{i(r)j(s)} - E_{i(t)j(s)}] > 0 \qquad (2)$$

In words: If we take the set of rotamers j(s) for each other residue j which give the minimum difference in energy between i(r) and i(t) (i.e. again we are trying to favour i(r)), and sum the differences in interaction energy, and the result is greater than 0[i] , then on average, $E_{i(t)j(s)}$ is lower in energy than $E_{i(r)j(s)}$ and we can therefore say that i(r) is dead-ending.

Criterion (2) is less stringent, and more effective in elimination, than (1). The energy difference $E_{i(r)j(s)} - E_{i(t)j(s)}$ will be greater (more positive) than the difference between min $E_{i(r)j(s)}$ and max $E_{i(t)j(s)}$ in (1), because in (1) we can choose two rotamers of j to favour i(r) and disfavour i(t), and in (2) we can only choose the one rotamer to do this. Therefore i(r) is less likely to be favoured and more likely to be eliminated.

Typically, therefore, criterion (2) would be applied to eliminate rotamers. The fact that the set of rotamers available would change (due to elimination) means that a second

i We also need to take into account the self energies $E_{i(r)}$ and $E_{i(t)}$, as the criterion shows

iteration of criterion (2) could be performed with different results,and more rotamers could be eliminated. Iterations would then continue until no more rotamers are removed.

As a way of eliminating yet more rotamers, the above criteria can be extended to pairs of rotamers, that is to find out

.

whether a rotamer pair i(r)j(s) is dead ending with respect to another pair i(t)j(u). To achieve this the interactions with the other residues k are considered in a similar way to single dead-ending rotamers, yielding criteria (3) and (4), corresponding to (1) and (2):

$$E_{i(r)j(s)} + \Sigma_k \min_v E_{[i(r)j(s)]k(v)} > E_{i(t)j(u)} + \Sigma_k \max_v E_{[i(t)j(u)]k(v)} \quad (3)$$

$$E_{i(r)j(s)} - E_{i(t)j(u)} + \Sigma_k \min_v \{E_{[i(r)j[s]]k(v)} - E_{[i(t)j(u)]k(v)}\} > 0 \quad (4)$$

The dead-ending pairs would typically be used as follows. First single rotamers would be eliminated, then dead-ending pairs would be flagged. In the second iteration, any pairs i(r)j(s) in criterion (2) which are dead-ending pairs would not

be considered, which may lead to further elimination of rotamers. Then a further dead-end pair search would be done, and so on.

The algorithm as actually used can be summarised as follows (see Lasters et al., 1995 for full details):

i) Use criterion (2) for single dead-end elimination.
ii) Use criterion (3) to flag a number of dead-end pairs.
iii) Use criterion (4), which is more effective but slower than (3), to flag more dead-end pairs. Using (3) beforehand will speed up the process as any dead-end pairs flagged with (3) will not be considered.
iv) Iterate again, ignoring dead-end rotamers and pairs from the first iteration, until no more rotamers are eliminated.

The dead-end theorem has been shown to predict 70% of buried residues accurately for PTI (pancreatic trypsin inhibitor).

iii) Other methods

Wilson et al (1993) use a cluster search method, useful when sidechains on all loops of a structure are to be built. The sidechains to be modelled are built in random order, and using

a rotamer library, the lowest combination of rotamers for the sidechain and its surrounding sidechains is built. This is repeated for the next residue, which may involve changing some of the rotamers selected in the first sidechain. This is performed for all the sidechains and then another iteration is started, rebuilding each cluster to take into account rotamers selected in the previous iteration, and iteration is repeated until convergence. This has the advantage over the CONGEN method of not being biased to such a great extent by the first sidechain built (due to the clusters) and also includes a solvation term, unlike CONGEN. It would also converge more quickly than theoretically-based methods such as the dead-end elimination algorithm.

The Monte-Carlo algorithm has been used in a number of methods (Holm and Sander 1991, Lee and Subbiah 1991). First, some change is made to the system, such as altering torsion angles. If the energy is lowered, the new configuration is accepted, and if the energy increases, the new configuration is either accepted or rejected. Acceptance depends on a probability factor (the 'temperature'). This is reduced throughout the run, so that at first, higher energy configurations are frequently accepted, but later, they are more frequently rejected. This is a means of avoiding false minima in that the entire energy surface is at first explored, and then

the molecule is allowed to settle gradually into the true minimum. The main disadvantage with Monte-Carlo methods is becoming trapped in false minima.

Holm and Sander have used Monte-Carlo with rotamer libraries. A random rotamer from the library is placed at a random residue, and accepted if the van der Waals energy is reduced. If not, another random rotamer is placed, depending on the 'temperature' (see above). Lee and Subbiah also use Monte-Carlo, but in combination with conformational search; adjustments of +/- 10 degrees are made to a chi torsion of a sidechain (the initial conformation is random) and acceptance or rejection takes place in a similar way.

There are also knowledge based approaches to sidechain modelling. Levitt (1992) builds the backbone and sidechain using segments from known structures, and Laughton (1994) uses a 3-dimensional homology method, where the neighbouring residues of a sidechain to be built are estimated, and the closest match in a database of residues with sets of neighbours in terms of direction and distance from the C-beta is taken. Scoring is based on the Dayhoff mutation matrix and the distance of each corresponding neighbour in the database and model, after optimal superimposition. This method suffers from the disadvantage that the rarer residues, such as Trp, Cys,

His and Met, are not well represented in the database, leading to inaccurate models.

Most of the sidechain modelling procedures predict around 70% of chi-1 angles accurately, and the average sidechain RMSD is 1.5 - 2.0 Å. A recurring problem is that exposed sidechains are not predicted accurately, as their conformation is affected by water. Indeed, if only buried residues are considered, around 80% of chi-1 angles are predicted accurately, and the average sidechain RMSD is 1.0 - 1.5 Å. A further problem is that charged residues are frequently not predicted accurately, owing to the lack of electrostatic terms from most of the methods. Serine residues are also hard to predict, due to their small size which means that an incorrect conformation would not necessarily cause a van der Waals clash.

Accessibility scores

As seen in Chapter 2, when the CONGEN rebuild regions were changed, it has been observed, among 24 uncomplexed antibody CDR-H3 crystal structures ranging in length from 7 to 12 residues, that certain key residues are persistently exposed (>30% relative accessibility), whilst others are persistently buried (<30% relative accessibility) (see Table 1, page 65). These are summarised as follows (notation: N1=N terminal residue, N3=3rd from N terminal, C3=3rd from C terminal, etc.)

Length>	7	8	9	10	11	12
N1	**Buried**	**Buried**	**Buried**	**Buried**	**Buried**	**Buried**
N2	*Exposed*	-	-	-	-	-
N3	-	*Exposed*	*Exposed*	*Exposed*	*Exposed*	*Exposed*
N4	-	-	-	*Exposed*	*Exposed*	*Exposed*
N5	-	-	-	-	*Exposed*	*Exposed*
C4	-	**Buried**	**Buried**	**Buried**	**Buried**	**Buried**
C3	**Buried**	**Buried**	**Buried**	**Buried**	**Buried**	**Buried**
C2	-	-	-	-	-	-
C1	*Exposed*	*Exposed*	*Exposed*	*Exposed*	*Exposed*	*Exposed*

Given this pattern, a screen for generated conformations can be devised. Each conformation can be scored based on the amount that the accessibility of each key residue deviates from the mean observed accessibility in crystal structures, and then

divide by the standard deviation as a measure of the certainty of the observed accessibilities:

$$\text{Score} = \Sigma_{\text{key res}}(O-M)^2/s^f$$

where 'key res' are the key residues above, O is the observed accessibility for that residue in the model, M is the mean accessibility for that residue and that loop length in the crystal structures, s is the standard deviation and f is the 'standard deviation factor' (s.d.f.), a factor by which the standard deviation is raised to influence its effect on the score.

With this method, low scores mean good models in terms of forming the typical accessibility pattern, and vice-versa.

In order to use the method as a screen, the top five VFF energy models from the sidechain addition (either dead-end or CONGEN) from the previous investigation were scored using the above equation. The RMSD rankings were then compared to those using VFF.

5.3 Screening Methods

Sidechain placement

Models were generated using the revised modelling procedure. The non-H3 CDR sidechains were included in CONGEN and VFF; 1,4 interactions were ignored in VFF and the initial set of CONGEN-built H3 sidechains stripped off prior to the VFF minimisation (these were built on unminimised structures and had already been shown to lead to inter-sidechain clashes). RMSD clustering was performed on the conformations.

Two methods were employed to add the sidechains : the CONGEN Iterative algorithm, and the dead-end elimination algorithm. As noted before, this latter method is theoretically robust, finding the lowest energy configuration, and lacks the problems of some of the other sidechain placement methods, such as the false minima of the Monte-Carlo methods. The bottom 10 clustered conformations for each of the eight structures (except 1mam, in which all the models had low RMSD) and two new structures: 1kem (28B4; catalytic antibody Fab fragment; Hsieh-Wilson et al, 1996) and 1ucb

(BR96; anti-tumour; Sheriff et al, 1996) were taken, and sidechains added using the two methods.

The success of the two approaches were then compared as follows:.

i) To examine the accuracy of sidechain placement, the chi-1 and chi-2 angles for each buried (<30% relative accessibility) residue over all the structures were examined, and the percentages within 30 degrees of the crystal structure value for chi-1 only and both chi-1 and chi-2 were noted. The accuracies of dead-end and CONGEN were compared, both overall and for each individual residue type.

Note that buried residues only were used - in assessing the accuracy of a sidechain algorithm, exposed residues would present the additional problem of flexibility and are therefore justifiably left out at this stage.

ii) To examine the effectiveness of the methods as a screen, the RMSD distribution of the bottom 5 (by VFF) backbones and the bottom 5 (by VFF) conformations with sidechains added were compared. In addition, the bottom 5 conformations by VFF with sidechains added were ranked using the accessibility score method.

5.4 Results

Comparing the accuracy of prediction of chi angles in dead-end and CONGEN

Table 15 and Figure 18 show that for 4 out of these 9 residues (Arg,Asn,Phe and Tyr), both methods get the chi-1 correct 50% of the time; the residues for which chi-1 is not predicted accurately 50% of the time by either method are Glu,Asp and Ser. Only Phe and Asn, however, have both chi-1 and chi-2 correct as much as 50% of the time with dead-end, and with CONGEN the best performance for chi-1 and 2 prediction is Phe with 33%.

The most notable differences in performance between the two methods are that Met, Tyr and Asn are modelled with more success using dead-end, while Leu is modelled more accurately with CONGEN. Otherwise the two methods perform similarly.

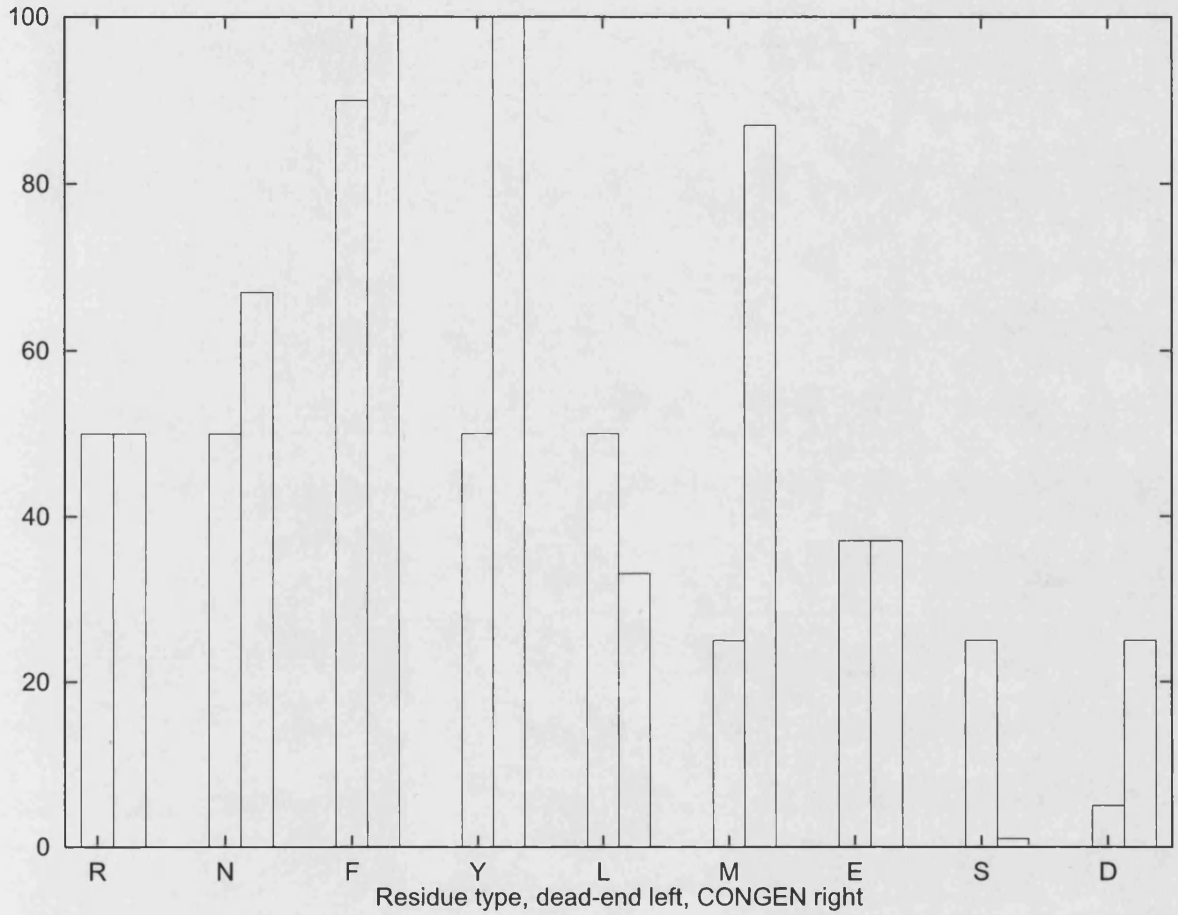


Figure 18. Percentage of chi-1 angles correct for various residue types when using the dead-end or CONGEN methods to construct the sidechains.

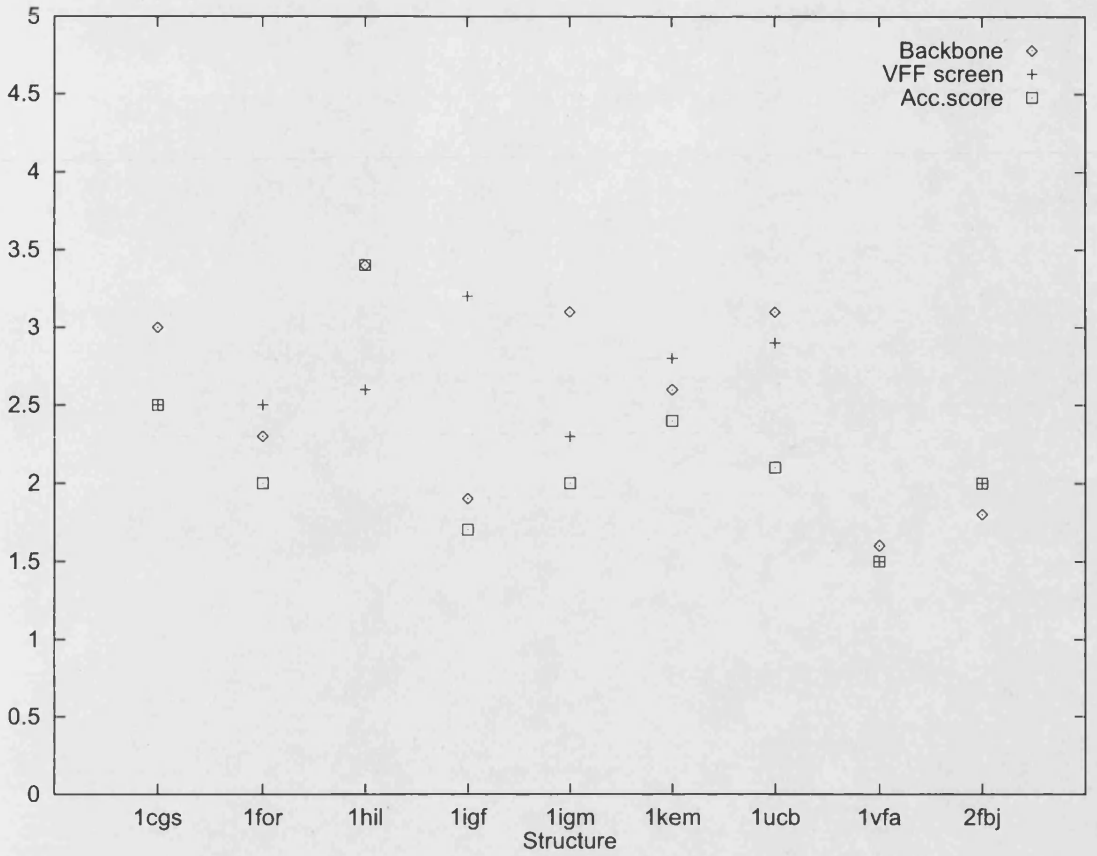
Table 15. The percentage of modelled sidechains for which chi angles are correct (+/- 30 degrees) in each residue type, in the bottom 4 (2 in 1hil) RMSD conformations in each modelled structure.

Residue type	%Chi-1 correct		%Chi-1&2 correct	
	CONGEN	Dead end	CONGEN	Dead end
Arg	50	50	30	0
Asn	50	67	17	67
Phe	90	100	40	60
Tyr	50	100	17	33
Leu	50	33	33	25
Met	25	87	0	38
Glu	37	37	38	0
Ser	25	0	25	0
Asp	5	25	5	5

The effect of adding sidechains, and the accessibility screen, on the accuracy of backbone prediction

The results (Table 16) show that:

i) adding sidechains to the backbones and then screening gives mixed results; for CONGEN, 1igm (notably), 1cgs, 1for and 1vfa show an improved RMSD spread, 1igf, 1vfa, 1hil, 2fbj and 1ucb show a similar spread, and 1kem a worse spread. For dead-end, the results are slightly worse: 1for shows no improvement and 1igf performs less well. So overall, there is a slight improvement for CONGEN and no improvement for dead-end. The RMSD of the very lowest energy conformation also shows no consistent improvement (Figure 19).



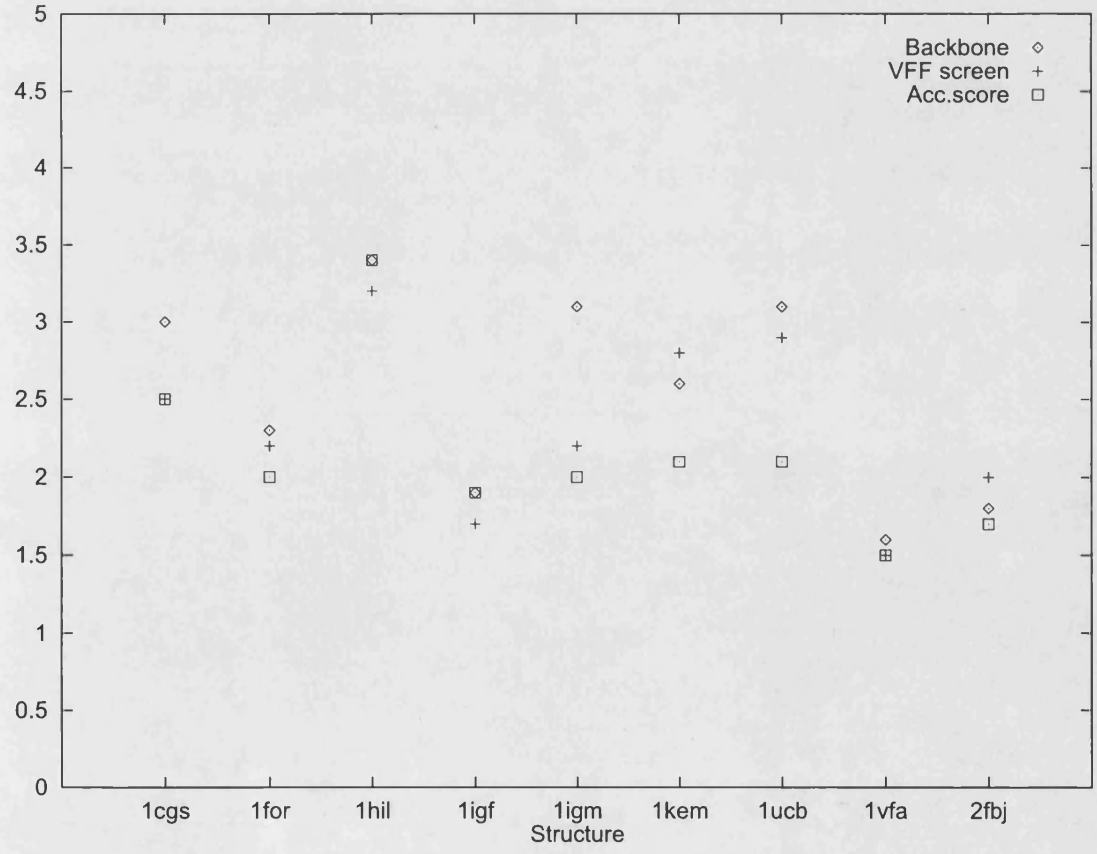


Figure 19. The RMSD of the lowest energy conformation using the 'new' modelling procedure with the following screens: VFF (backbone only), VFF (backbone and sidechains) and the accessibility score screen (backbone and sidechains). Sidechains were built with dead-end (top) or CONGEN (bottom).

ii) ordering the bottom 5 backbones (with sidechains added) using the accessibility screen gives an improved RMSD order compared to the VFF screen. In CONGEN, 1igf, 1igm, 1kem and 1ucb (4/9) improve, 1vfa, 2fbj, 1hil and1cgs (3/9) perform similarly, and 1for (1/9) does rather worse. In dead-end, 1igf, 1igm, 1kem, 1ucb and 1for (5/9) improve, 1vfa, 2fbj, 1cgs and 1hil (4/9) perform similarly. So overall, the improvement is notable, more so for dead-end than CONGEN.

In addition, all structures except 1hil select a conformation of RMSD <= 2.5 Å for both dead-end and CONGEN as the best scoring, while 6/9 structures (except 1hil, 1cgs and 1kem using dead-end) select a conformation of RMSD <= 2.2 Å. Furthermore, the RMSD of the structure with the very best accessibility score is lower than in 5/9 structures, and lower than or equal to in 7/9, that of the lowest energy conformation by VFF (Figure 19), using either dead-end or CONGEN to build the sidechains.

Table 16. The backbone global RMSD of the bottom 5 conformations by a number of screens;
'backbone only' - bottom 5 backbones after clustering the bottom 200;
'VFF screen' (dead-end and CONGEN) - add sidechains to the bottom 10 backbones after clustering the bottom 200, and then take the bottom 5 by VFF energy;
'Accessibility' - order the bottom 5 by VFF energy in terms of accessibility score. The sdf (see Method section) is 2.

Backbone only	Dead end VFF screen	Dead end Accessibility	CONGEN VFF screen	CONGEN Accessibility
1cgs				
3.0	2.5	2.5	2.5	2.5
2.3	3.0	3,1	3.0	3.1
3.0	3,1	3.0	3.0	3.0
3.0	3.0	3.0	3.0	3.0
2.5	3.0	3.0	3.1	3.0
1for				
2.3	2.5	2.0	2.2	2.0
2.2	2.2	2.2	1.3	2.2
2.5	2.0	2.5	2.5	2,5
2,1	2.1	2.1	2.0	1.3
2.5	2.8	2.8	2.1	2.1
1hil				
3.4	2.6	3.4	3.2	3.4
3.3	3.3	3.3	2.6	3.9
2.2	3.9	3.9	3.9	2.9
4.8	3.4	2.6	2.9	2.6
3,4	4.8	4.8	3.4	3.2
1igf				
1.9	3.2	1.7	1.7	1.9
1.9	3.4	1.9	1.9	1,7
3.4	1.9	3.0	3.2	1.4
2.3	1.7	3.2	1.4	3.4
1,.8	3.0	3.4	3.4	3.2
1igm				
3.1	2.3	2.0	2.2	2.0
2.3	2.3	2.3	2.3	2.3
2.0	2.2	2.1	2.1	2.1
3.6	2.0	2.3	2.3	2.3
2.3	2.1	2.2	2.0	2.2
1kem				
2.6	2.8	2.4	2.8	2.1
2.4	3.7	2.6	3.0	1.8
3.0	2.2	2.2	3.7	2.8
2.9	2.4	2.8	1.8	3.0
1.8	2.6	3.7	2.1	3.7
1ucb				
3.1	2.9	2.1	2.9	2.1
2.5	2.1	2.4	2.1	2.9
2.9	2.4	2.9	3.1	2.6
3.3	3.1	3.3	3.3	3.3
2.6	3.3	3.1	2.6	3.1

1vfa				
1.6	1.5	1.5	1.5	1.5
3.3	1.5	1.5	1.9	1.9
1.5	2.2	2.2	2.4	2.2
2.4	2.4	2.4	2.2	1.5
1.8	3.3	3.3	1.5	2.4
2fbj				
1.8	2.0	2.0	2.0	1.7
2.0	1.7	1.0	1.7	1.0
1.0	2.3	2.3	2.3	2,3
2.5	1.0	1.7	1.0	2.0
1.7	2.5	2.5	2.5	2.5

5.5 Conclusions

It has been seen that the two alterations to the modelling procedure, that is, removal of intra-loop 1,4 interactions and inclusion of the sidechains of the other CDRs in both CONGEN and VFF minimisation while H3 sidechains are deleted, leads to improved results. Both improvements are needed: neither give consistently improved results on their own. The improvement that inclusion of the other CDR sidechains produces shows that an accurate environment when building the H3 loops is necessary, with all sidechains present. In this case the other sidechains were from the crystal structure; in an unknown this would not be so. How this problem can addressed will be discussed later.

Accuracy of modelled sidechains

The two methods that have been employed to add sidechains to minimised structures, the dead-end elimination algorithm and CONGEN, have varying degrees of success in predicting chi-1 and chi-2 angles, depending on the residue type (see Results section). The residue types for which chi-1 is predicted correctly most frequently are residues with large sidechains, namely Arg, Phe and Tyr; these together with Asn

have chi-1 predicted correctly 50% of the time in both dead-end and CONGEN. This is likely to be because a change in chi-1 would lead to a change in the overall direction of the sidechain, leading to a steric clash. Conversely, the residues for which chi-1 is predicted correctly *less* than 50% of the time with *both* dead-end *and* CONGEN are Glu,Asp and Ser. For the first two types, an inaccurate chi-1 could lead to an energetically acceptable structure due to formation of a 'false' salt-bridge, and Ser is so small that an inaccurate chi-1 would not necessarily lead to a steric clash. Indeed, Ser is the poorest modelled of all.

The chi-2 angle is predicted accurately much less frequently than chi-1; the only residue for which it is predicted accurately 33% of the time using both methods is Phe. This appears to indicate that getting the overall direction of the sidechain (determined by chi-1) is the most important factor, and that there is still some degree of flexibility remaining for chi-2 once the overall direction has been determined. Indeed the residues for which there is the biggest difference in performance between accurate chi-1 and accurate chi-1 and 2 prediction are Phe and Tyr; for these types, altering chi-2 would merely lead to rotation of the ring, which does not at all change the direction of the sidechain.

Lastly in the discussion of chi angle prediction, certain residue types are predicted more accurately with dead-end and others with CONGEN, i.e. Met, Tyr and Asn with dead-end and Leu with CONGEN. This does not show any patterns with regard to residue type and indeed the small data set (7 structures; four conformations for 6 of the 7 and two for the other structure) means we cannot be sure this is an absolute pattern. Testing on more structures as they become available would be useful here; if a pattern emerges, it could be sensible to employ a sidechain placement algorithm which used CONGEN for certain residue types and dead-end for others. A problem becomes apparent here, however, in that one method would have to be employed first, and while performing that method, sidechains of the CDR to be built with the other method would have to be ignored.

Filtering methods

Sidechain placement as a filter for inaccurate conformations does not appear to be that effective, with a similar RMSD spread in the bottom 5 energy (by VFF) conformations with and without sidechains. This indicates that the inaccurate conformations can still accommodate

sidechains without steric clashes resulting. What is more effective, though, after sidechain addition, is the accessibility screen on the bottom 5 conformations by VFF, which in the majority of structures successfully filters out the very high RMSD conformations (>3.0Å) (Figure 20), and in a number of structures, notably 1vfa, 1igf and 1igm, is particularly effective at selecting the lowest RMSD conformations available in the bottom 5 conformations by VFF. All the structures out of the 9 examined select a conformation of RMSD 2.5Å or less (Figures 21a, 21b) except 1hil. In this instance, there was no conformation of RMSD 2.5Å or less in the bottom 5 conformations by VFF. However, there are 2 conformations in this RMSD range in the bottom 10, one below 2Å, and the accessibility screen still fails to select either of these 2 (unpublished data) when the bottom 10 are examined. The loop of 1hil is a long loop (11 residues) and may exhibit greater flexibility in solution than other H3 loops of similar length.

In summary, the sidechain build + VFF + accessibility scoring method presented here is an effective screen for H3 backbones of up to 10 residues, in general, with a reasonable conformation in the final 5 in almost every case (Figures 21a, 21b) and in some instances (e.g. 1vfa, 2fbj, 1for, 1igm) several reasonable conformations are present in the final 5 (Figures

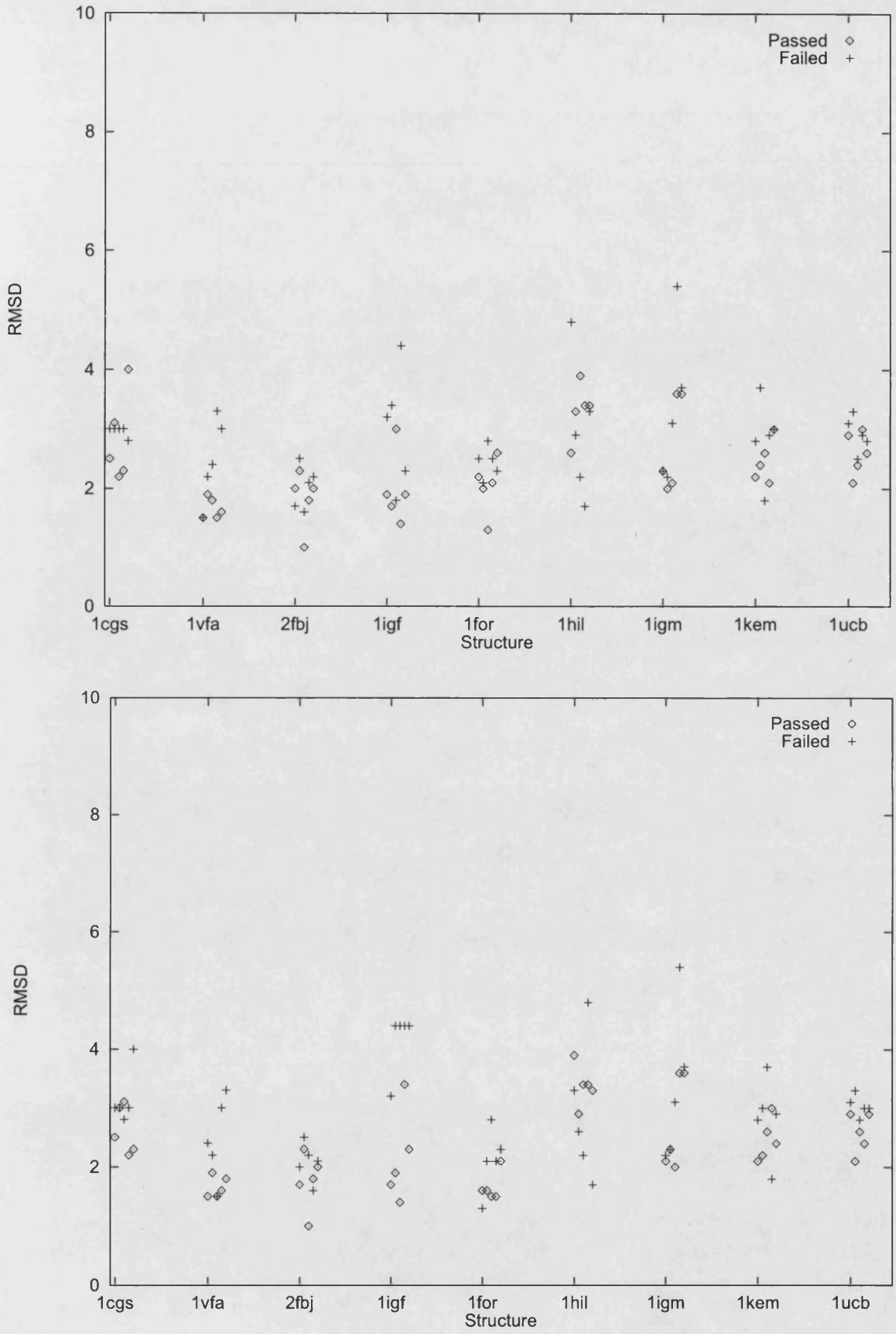


Figure 20. The effectiveness of the accessibility screen.
The bottom 10 backbones by VFF, with sidechains built by (top) dead-end, (bottom) CONGEN, were screened with the accessibility score method,, the best 5 being the 'passes' and the others being 'failures'. The plot shows that for most structures, with the exception of the 11-residue H3 of 1hil, the method is a good screen for lower RMSD conformations.

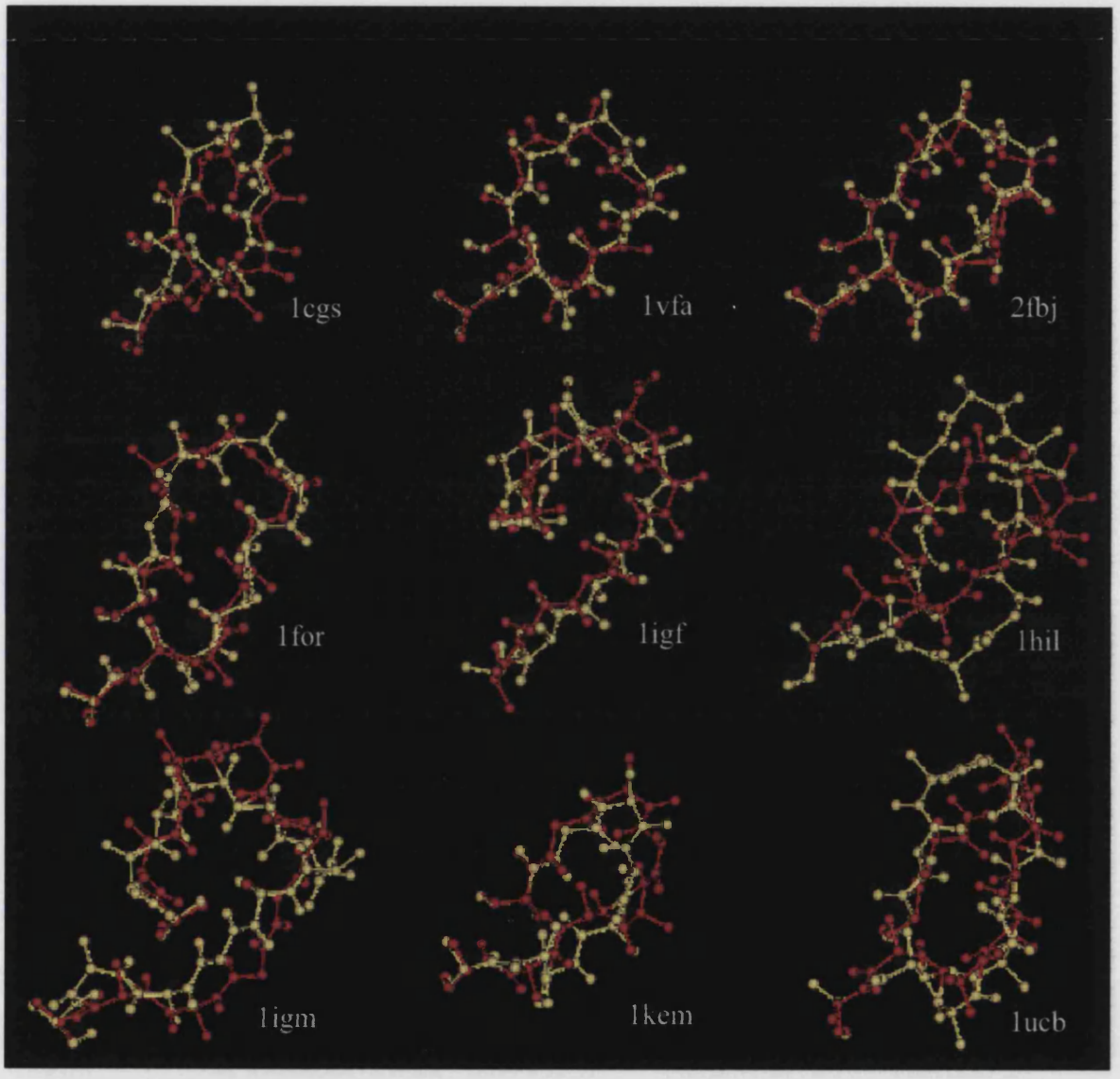


Figure 21a. The lowest RMSD CDR-H3 model backbone of the final five models (red) superimposed on the crystal structure backbone (yellow) for all the structures, using CONGEN to build the sidechains and the accessibility score method to select the five final conformations.

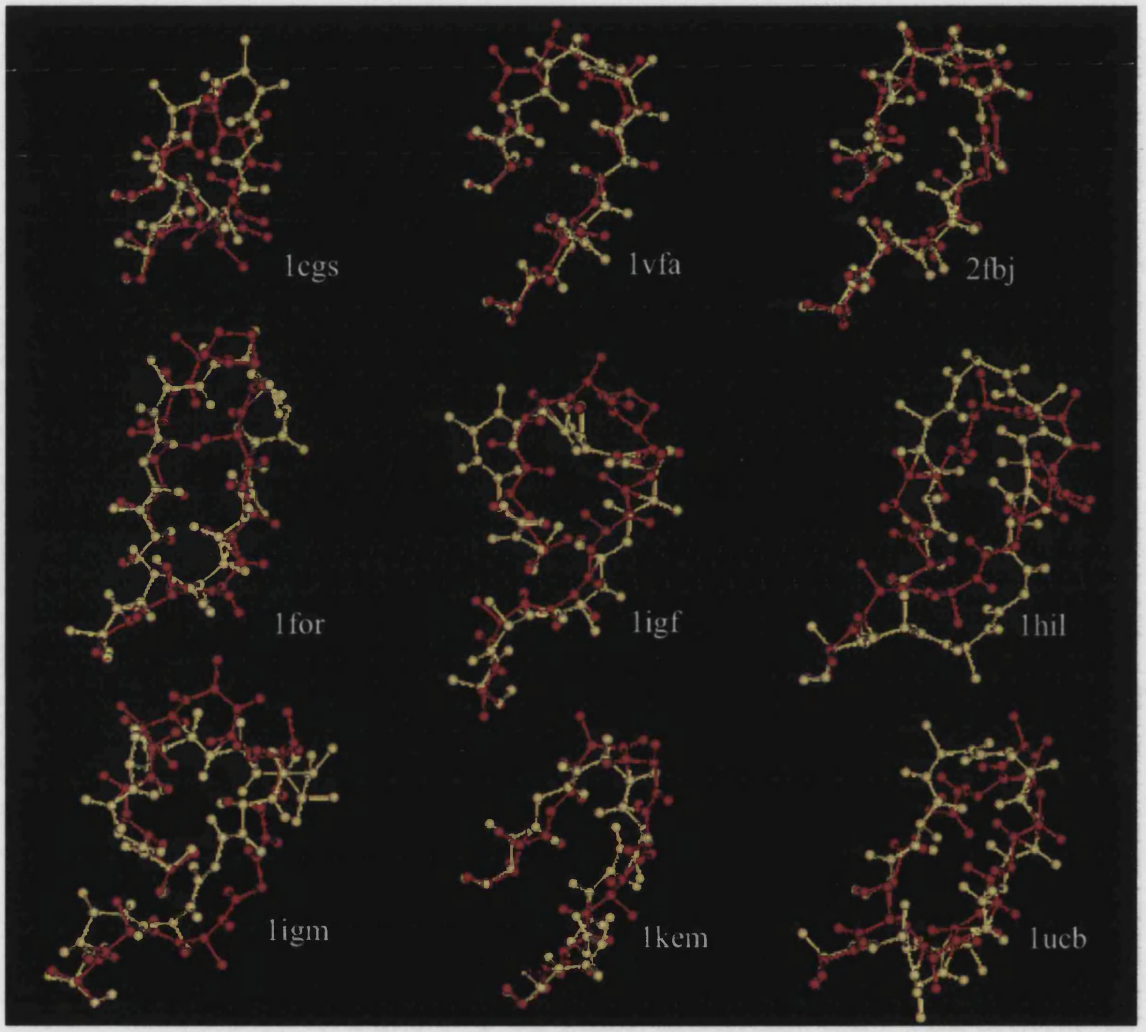


Figure 21b. The lowest RMSD CDR-H3 model backbone of the final five models (red) superimposed on the crystal structure backbone (yellow) for all the structures, using the dead-end method to build the sidechains and the accessibility score method to select the five final conformations.

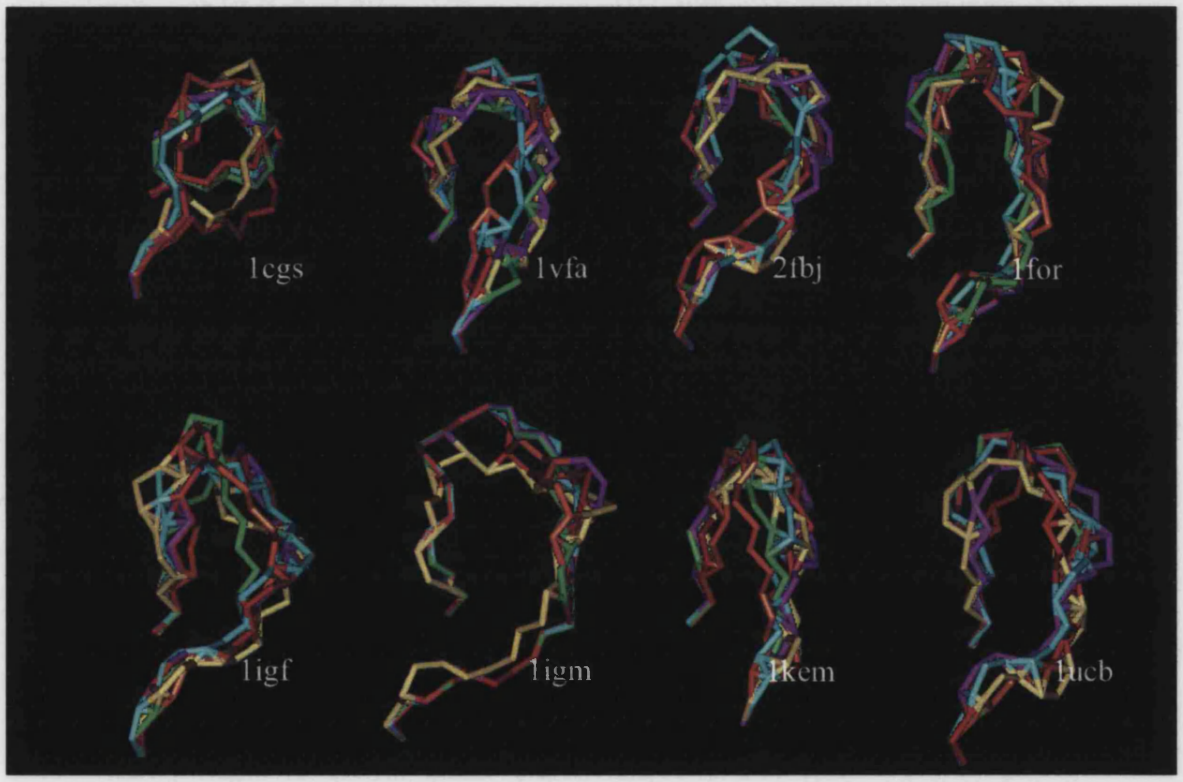


Figure 21c. The five final model backbones (selected by the accessibility score method) superimposed on the crystal structure backbone (yellow) for all the structures (except the poorly modelled 1hil), using CONGEN to build the sidechains.

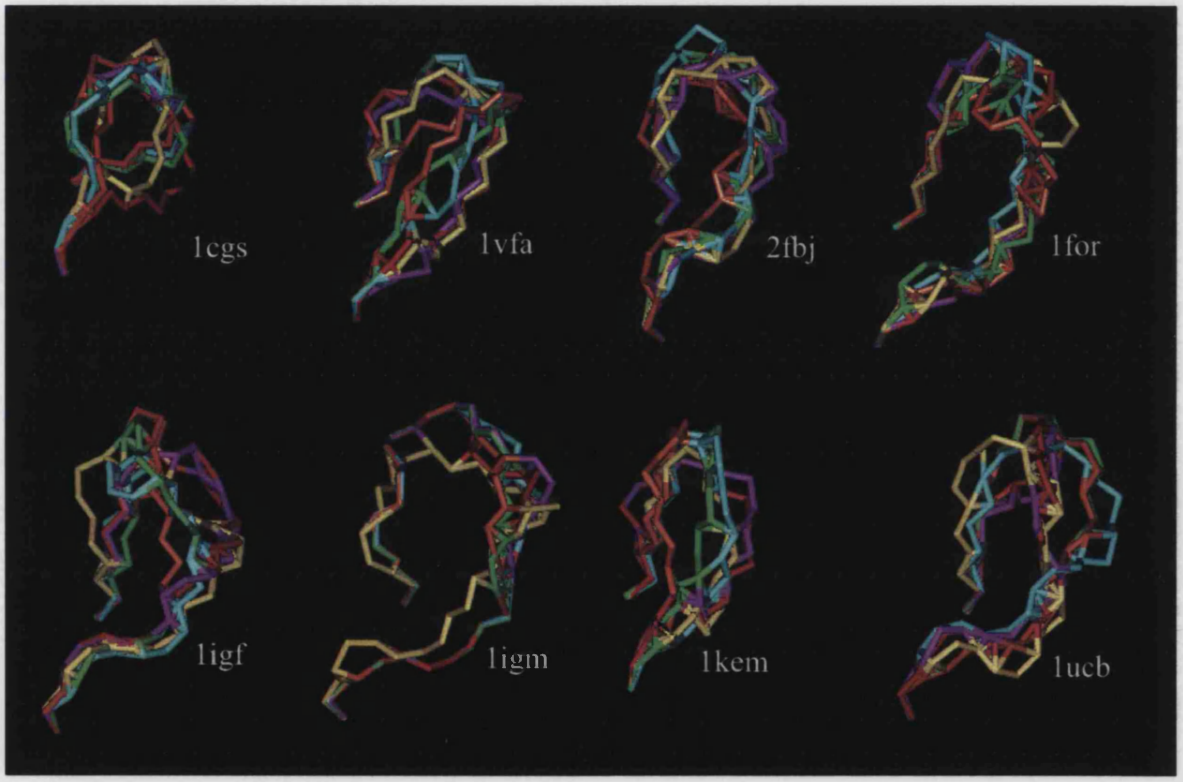


Figure 21d. The five final model backbones (selected by the accessibility score method) superimposed on the crystal structure backbone (yellow) for all the structures (except the poorly modelled 1hil) using the dead-end method to build the sidechains.

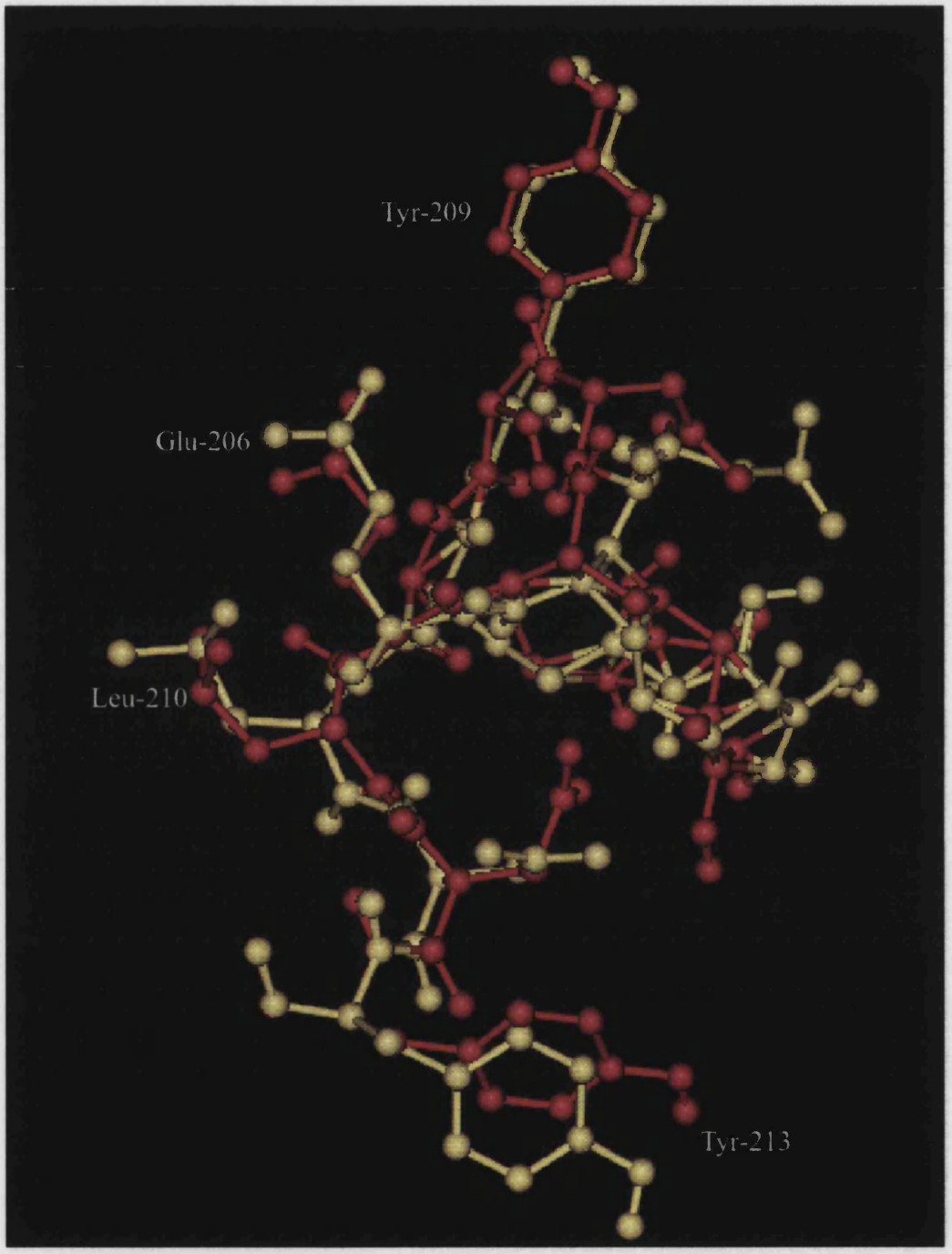


Figure 21e. The lowest RMSD CDR-H3 model (including sidechains) of the five final models (red) superimposed on the crystal structure H3 (yellow), for the best-modelled structure (1vfa), using CONGEN to build the sidechains and the accessibility score method to select the five final conformations.

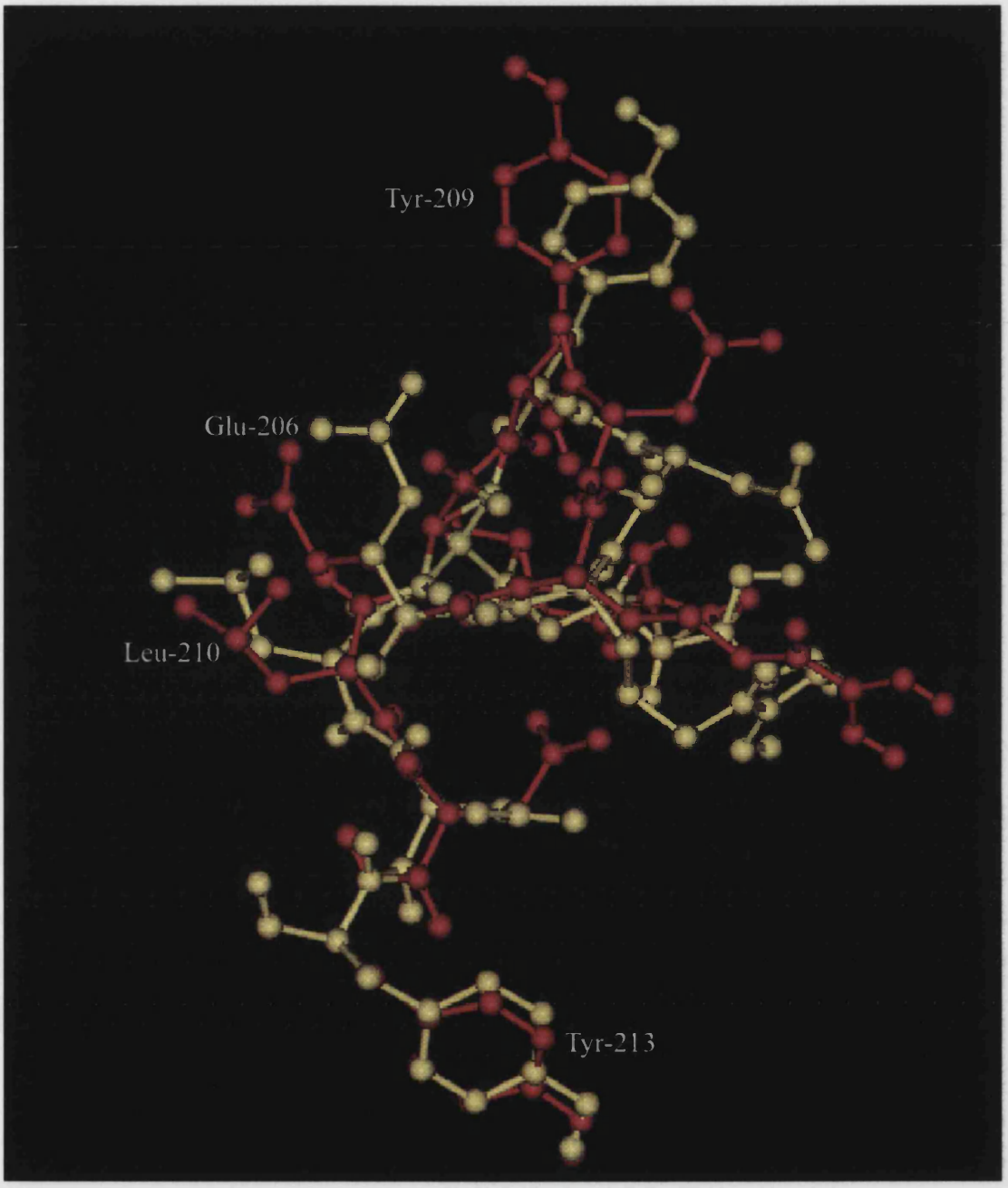


Figure 21f. The lowest RMSD H3 model (including sidechains) of the five final models (red) superimposed on the crystal structure H3 (yellow), for the best-modelled structure (1vfa), using the dead-end method to build the sidechains and the accessibility score method to select the five final conformations.

21c, 21d). As seen above, the accuracy of the sidechains is dependent on residue type, but aromatic residues in particular are predicted accurately (Figures 21e, 21f).

One crystal structure loop, 1mam, does not have the typical shape of an H3 loop (see Method) and consequently does not show the typical accessibility pattern in the crystal structure. In this instance, it would be useful to ensure that there are no features (such as the two prolines in 1mam) of the sequence of the H3 loop that are likely to distort the normal 'kinked' shape to the 'extended' shape (Chapter 6; Shirai et al, 1996). For the longer loops, it may be necessary to resort to molecular dynamics, running a simulation of the loop (together with the rest of the Fv) in water to get a set of possible conformations, and as a test, compare the median of these conformations to the crystal structure for a known structure.

The accessibility score method appears to be more effective as a way of taking into account the solvation effect than the solvation energy method of Eisenberg and McLachlan (1986), as the score method takes into account the consistent pattern observed in antibody H3 loops, whereas the Eisenberg and McLachlan method assumes that hydrophobic residues will always prefer to be buried and vice-versa. This is not necessarily so: charged or hydrophilic residues may well be buried forming salt bridges (such as that formed in an H3 loop

between residues 219 and 235 when these are Arg and Asp respectively) or hydrogen bonds, and conversely, hydrophobic residues, particularly Phe and Tyr, may be exposed in order to form contacts with antigen.

CHAPTER 6 - A CANONICAL FEATURE FOR H3

6.1 Introduction

As seen in Chapter 1, five of the six CDRs fall into a number of canonical classes. CDR-H3, however, is more variable in structure.

Now that more structures are available in the PDB, there have been several investigations to try and define a set of rules for determining H3 structure. An early investigation was by Rees et al (1996), who defined a set of N-terminus/C-terminus interactions in certain 8-12 residue loops and in addition, attempted to classify H3 take-off angles as a function of length. The following sub-classes were defined.

H3a - 6 residues or less

H3b - 7 residues

H3c-f - loops of 8, 9/10, 11 or 12 residues, in which a conserved Arg or Lys at position 219 interacts with Asp, Gly or Ala at position 235.

Longer loops did not easily fit into the classification and so were not categorised.

Further investigations have since been performed, such as that by Shirai et al (1996). They observed that the C-terminal conformation fell into two groups; 'kinked', in which residue 234 points inwards and 235 outwards and 'extended' in which a standard beta-strand extended conformation is assumed, with residue 234 pointing

outwards and 235 inwards, in contrast to the 'kinked' group. The majority of structures fell into the 'kinked' group.

A number of rules were formulated for deciding which conformation would be adopted (see Figure 22):

a) If residue 235 is **not** Asp, a kinked structure is formed due to a hydrogen bond between the carbonyl O of residue 234 and the ring N of Trp-237.

b) If residue 235 **is** Asp, but residue 219 is **not** Arg or Lys, the carboxyl of Asp-235, rather than the carbonyl of residue 234, forms the hydrogen bond with Trp-237, forming the extended structure.

c) If residue 235 is Asp, **and** residue 219 is Arg or Lys, the carboxyl of Asp-235 forms a salt bridge with residue 219 rather than a hydrogen bond with residue 237. This, together with the 234-237 interaction in a), forms the kinked structure.

d) If residue 235 is Asp, and both residue 219 **and 218** are Arg or Lys, the carboxyl of Asp-235 forms a salt bridge with residue 218 rather than 219, and the extended structure is formed.

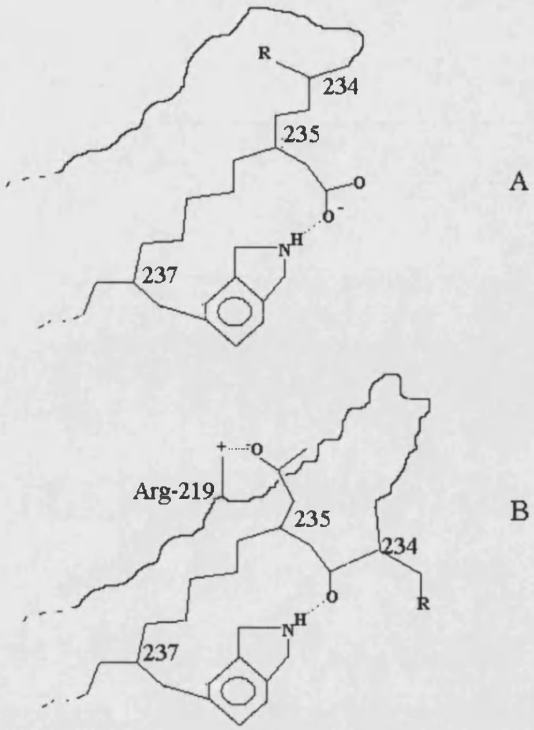


Figure 22. The kinked and extended structures.

A:The extended structure. Asp 235 points into the interior and H bonds with Trp-237. The sidechain of residue 234 points outwards.

B: The kinked structure. Asp-235 points outwards and H bonds with Arg-219. Trp-237 H bonds with the carbonyl of residue 234. The sidechain of residue 234 points inwards, into the pocket.

The majority of structures conformed to these rules. There were two exceptions: 1mam and 1igi. The former has an Ala rather than Asp at residue 235, and so would be expected to form the kinked structure, according to a). However, it is extended. This is because there are two prolines in the 8-residue loop, which distort the structure. In 1igi, residue 235 is Asp and residue 219 is not Arg or Lys, so the extended structure would be expected. However, the carboxyl of Asp-235 forms a salt bridge with a lysine within the H3 loop, rather than with Trp-237, and the kinked structure is thus formed by the 234-237 interaction in a).

The authors also postulated a number of additional rules for H3 structure:

a) 'Beta bulge' formation in kinked structures. In most longer H3 loops, a beta-ladder (of varying sharpness of definition) is formed in the region of the H3 between the head of the loop and the terminals, with residue 234-1[i] being the C-terminal end of the ladder. If, however, residue 234 is glycine and 233 is large and hydrophobic, residue 234-1 projects into the

i The residue preceding 234 is numbered differently according to the number of deletions in the loop; so it is denoted 234-1.

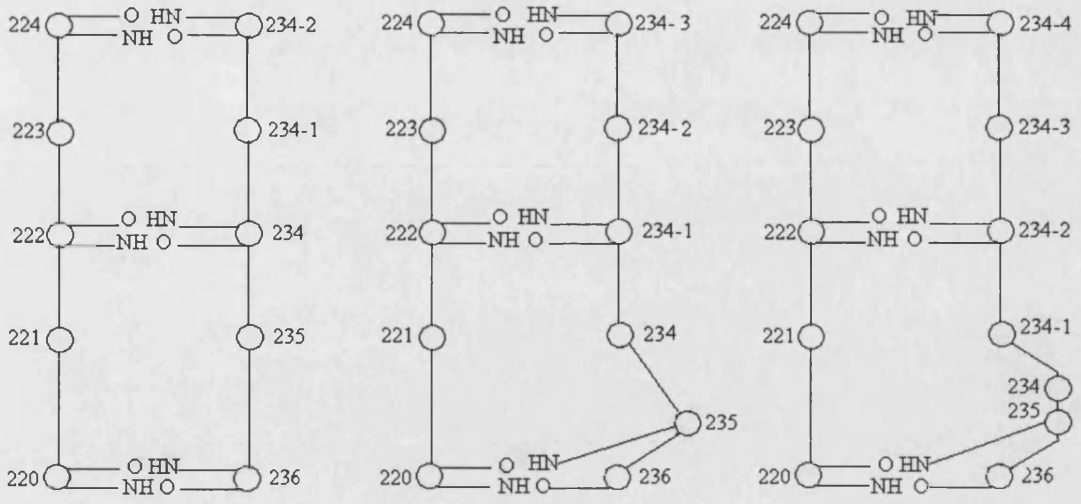


Figure 23. The differing patterns of peptide hydrogen bonding in the extended (left), kinked (middle) and kinked with beta bulge (right) classes of H3 loops.

hydrophobic interior, rather than residue 234 which normally does so, leading to a 'beta bulge' and altering the hydrogen-bonding residues of the beta-ladder (see Figure 23).
This also takes place if residue 234-1 is Trp (irrespective of what residue 234 is) as there is too much steric hindrance (from residue 235 and the light chain) to allow it to be placed on the outside.

b) If the first, or second, residues on both the N and C terminal side of the beta-ladder region, are aromatic, the strands are drawn together, sharpening the definition of the beta-ladder and forming a typical hairpin structure. However, if there is a Pro at one of the positions where a hydrogen bond pair across the loop is formed, the ladder is disrupted.

c) For the hairpin-forming structures in b), the head of the loop forms a typical beta-turn, with the subtype dependent on the positioning of Gly/Asp/Asn (which are more likely to form positive phi/psi angles) or Pro (which can more easily form a cis-peptide bond) in the turn (Sibanda and Thornton, 1993). However, for non hairpin-forming structures, the head of the loop is much less well defined.

Morea et al (1998) have also postulated sub-types for H3 loops. Their conclusions were similar to those of Shirai et al in that they subdivided the loops into 'torso bulged' and 'torso non bulged'

classes, corresponding to the kinked and extended classes. (The 'torso' refers to the terminal regions of the loop). The rules drawn up for the terminal region conformation are generally the same as those of Shirai et al.; however, they did not mention the effect of residue 218 being Arg or Lys, and subdivided the 'torso non bulged' class into two subtypes, determined by interactions between H3 and the two conserved aromatic residues 143 and 151 of H1. One subtype is formed if there is an aromatic at residue 221 of H3, and the other if it is at residue 224.

Three exceptions to the rules were found; two, 1mam and 1igi, were exceptions in the work of Shirai et al., and the same reasons are postulated. A third, 1hil, is claimed not to form the 'torso bulged' structure (its expected structure). However, the basis of this claim is unclear, as examination with Insight II reveals a 'torso bulged' structure (kinked C-terminal) and furthermore the structure is placed in the 'kinked' class by Shirai et al.

Finally, Morea et al also conclude that the determinants of the head region of the loop are those for beta-turns.

6.2 Methods

Preliminary investigations

It has been shown (see Table 1, Chapter 2) that the H3 sidechain of residue 234 (third from the H3 C terminus; see Appendix 3 for numbering scheme) is frequently predicted accurately when the sidechain is rebuilt using either dead-end or CONGEN.

This suggests the possibility that this residue is located in an environment that is conserved from one antibody to another. Further analysis confirmed that residue 234 actually protrudes into a pocket, forming a number of contacts with framework residues, in the manner of canonical key residues (Chothia and Lesk 1987). Examination of the structures confirmed this, with 21/24 of the uncomplexed CDR-H3 crystal structures forming contacts (within 5Å) between the sidechains of residue 234 and the following framework residues:

Residue type	*Number*	*Position w.r.t. CDRs*
Tyr	42	2 after L1 C-terminus
Phe	106	1 after L3 C-terminus
Val	156	2 after H1 C-terminus
Trp	166	3 before H2 N-terminus

Further examination showed that in the 21/24 structures, the sidechain of residue 234 did indeed protrude into the barrel in a pocket, and was surrounded by the above framework residues. The

sidechain of Tyr-42 contacted the C-beta region of the sidechain of residue 234, whereas the sidechains of Phe-106, Trp-166 and Val-156 surrounded the end of the sidechain (see Figure 24).

Of the other crystal structures, 1mfb had the same general shape H3 loop as the 21/24 group, with the C-alpha of residue 234 pointing inwards (a common feature of the 21/24 group). However, in 1mfb residue 234 was a glycine forming no contacts and therefore leaving a 'hole'. The relative positions of the pocket sidechains were retained however. The other two antibodies, 1mam and 1mrc, had completely different shapes of H3 loops, with the C-alpha of residue 234, and thus the sidechain, pointing outwards, so that the contacts could not be made. These observations support those made above by Shirai et al (1996), with 1mam and 1mrc belonging to the extended class (which exposes residue 234), whereas the others belong to the kinked class (which buries it in the interior). As seen in the Introduction, the sequence of the H3 loop of 1mam (DPYGPAAY) has two prolines, distorting the structure, and forcing it to be 'extended'. The feature giving rise to the different structure in 1mrc (LRGYFDY) is the presence of

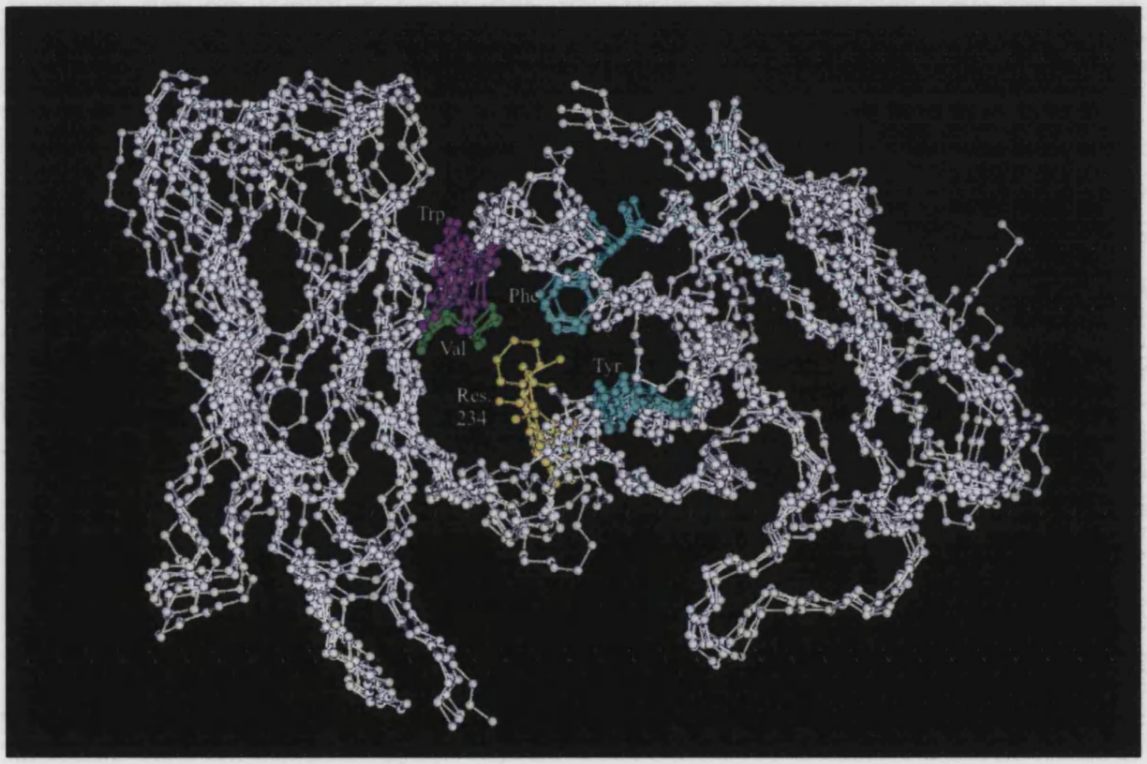


Figure 24a. The CDR-H3 hydrophobic pocket, as viewed from above.

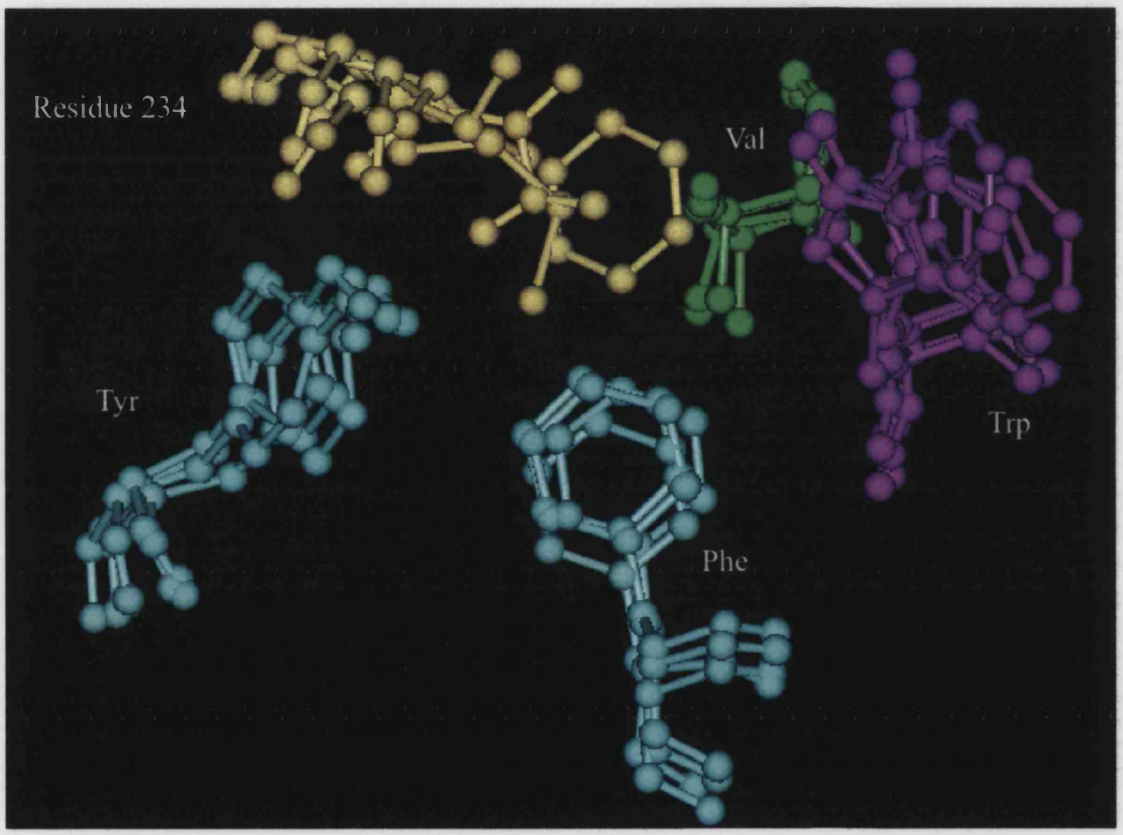


Figure 24b. CDR-H3 residue 234 and the residues of the hydrophobic pocket.

Asp at residue 235 but the absence of Arg or Lys at position 219 (see Introduction).

The observation of this 'canonical' feature in H3 loops suggests that it can be used as a screen for the generated conformations. Those conformations not forming the four contacts are rejected, for structures which have the correct H3 C-terminal structure to form the pocket, i.e. belong to Shirai's kinked class. This will now be investigated.

Using the contacts formed by residue 234 as a screen

The entire set of clustered backbones from the revised (see Chapter 5 introduction) modelling procedure (rather than the bottom 10; this was in order to give a larger conformation set) was taken, sidechains added with either dead-end or CONGEN and conformations for which less than 3 of the 4 contacts were made (sidechain to sidechain, within 5Å) were screened out. The remaining conformations were ranked using either VFF or the accessibility score.

6.3 Results

Tables 17 and 17a show that the contacting residues screen does not lead to an improvement for the majority of structures,

but for the structure (1hil) that continued to do badly in the accessibility results (Table 3), a dramatic improvement is seen, notably for CONGEN where the only conformation which survives the contacts screen is the sole conformation with rms < 2Å, using both VFF and the accessibility score. For dead-end, one other conformation (rms 4.0 Å) survives the screen; unfortunately, this one comes top using both VFF and the accessibility score.

Table 17. The backbone global RMSD of the bottom 5 conformations of the entire set of the clustered bottom 200 backbones, after sidechain addition, and before and after the contacts screen. NB: where there are less than 5 conformations, less than 5 survived the contacts screen.

Dead end	Dead end + contact screen	CONGEN	CONGEN + contact screen
1cgs			
2.5	2.5	2.5	2.5
3.0	2.8	2.8	2,8
4.4	3.1	4.4	3.1
2.8		3.1	
3.1		3.0	
1vfa			
1.8	1.5	1.5	1.9
2.1	2.1	1.9	2.3
1.5	1.9	3.4	2.1
1.6	2.2	2.4	2.2
1.5		2.2	
1igf			
3.2	2.9	2.9	2.9
2.9	1.9	1.7	1.8
3.4	1.7	1.9	1.8
1.9	3.0	3.2	
1.7	2,3	1.8	
1hil			
2.4	4.0	3.3	1.7
2.6	1.7	2.6	
3.3		2.4	
4.1		4.0	
3.2		2.6	
1igm			
3.0	2.3	2.2	2.2
3.5	2.3	2.3	2.3
2.3	2.2	2.1	2.1
2.3	2.6	2.3	2.3
2.2	2,.1	2.1	2.1
1for			
2.5	2.5	1.6	2.2
1.6	2.2	2.2	2.5
2.2	2.2	1.3	2.5
2.2	2.4	2.5	2.0
1.7	2.3	2.5	1.8
1ucb			
2.2	2.2	2.9	2.1
2.9	2.1	2.1	2.2
2.1	2.4	2.2	
5.2	1.5	5.2	
2.4		3.1	
2fbj			
2.0	2.0	2.0	2.0
2.3	2.3	1.7	1.7
2.4	2.4	2.5	2.5
1.7	1.7	2.4	2.4
2.3	2.3	3.3	2.3

Table 17a. The backbone global RMSD of the bottom 5 conformations (by accessibility score; sdf 2) of the entire set of the clustered bottom 200 backbones, after sidechain addition, and before and after the contacts screen. N.B: where there are less than 5 conformations, less than 5 survived the contacts screen.

Dead end	Dead end + contact screen	CONGEN	CONGEN + contact screen
1cgs			
2.1	3.1	3.1	3.1
2.2	2.5	2.8	2.8
3.1	2.8	1.8	2.5
2.5		2.1	
2.6		2.3	
1vfa			
1.9	1.9	1.9	1.9
2.1	2.1	2.1	2.1
2.0	1.5	2.2	2.2
2.4	2.2	1.6	2.3
1.9		1.5	
1igf			
1.7	1.7	1.8	1.8
1.7	1.9	1.9	1.8
1.9	1.9	1.8	2.9
1.9	1.8	1.7	
1.8	3.0	1.7	
1hil			
3.3	4.0	3.4	1.7
2.6	1.7	4.0	
4.0		4.0	
3.4		3.4	
3.2		4.7	
1igm			
3.6	3.6	2.6	2.6
2.4	2.4	3.6	3.6
2.0	2.0	2.1	2.1
3.0	3.0	2.4	2.4
2.6	2.6	3.6	3.6
1for			
2.0	2.1	2.1	2.1
1.3	2.0	1.7	1.8
1.7	2.2	1.8	2.3
1.6	1.7	1.4	1.8
2.1	1.8	2.5	2.4
1ucb			
1.5	1.5	1.5	2.2
2.2	2.2	2.2	2.1
2.1	2.1	2.4	
2.4	2.4	2.1	
2.9		2.9	
2fbj			
1.7	1.7	1.7	1.7
2.0	2.0	1.7	1.7
2.0	2.0	1.0	1.0
2.0	2.0	2.0	2.0
1.0		2.3	2.3

The only other improvement is 1ucb, with VFF but not with accessibility. 1cgs does worse in the accessibility screen.

6.4 Conclusions

The screen for the presence of three of the observed four framework contacts of residue 234 does not, in general, lead to improved results, either using VFF or the accessibility screen. It is evident from the results that a number of high rms structures still form these contacts. The only structure for which an improvement is seen, and it is a notable improvement, is 1hil, where (using CONGEN) all the high rms structures are filtered out, leaving the one structure with an rms of below 2.0Å. The reason for this observation could be again due to the fact that it is a long loop: a misplaced framework, and therefore sidechain, for residue 234 is more likely to be on the outside of the Fv, and therefore not in the cavity forming the contacts. This raises the question of whether this would be a good screening method for long loops; it is not effective for the other H3 with >10 residues, 1igm, but may be worth testing as more structures become available.

It is of note that residue 234, and the framework residue Phe-L106, are derived from the J genes of the heavy and light chain respectively, of which there are four subtypes in the mouse; it could be that the local structure around the C terminus of the H3 loop is determined by the subtype of J gene. This could be investigated further.

The observation of the 'canonical' pocket formed by residue 234, as well as the observations by Shirai et al. and Morea et al. (neither of which discovered the clear pattern of the four contacts mentioned here) does hint at the possibility of canonical structures for the H3 loop. At present, rules exist for the loop base, so a possible modelling method would involve using the base of the most homologous known structure in terms of sequence and loop length, and which belongs to the same class (kinked or extended) as the structure to be predicted. However, the head is less well defined, except in some circumstances where a well defined hairpin is formed (Shirai et al.,1996; see Introduction to this Chapter). At present, therefore, the existing methods of database or conformational search would have to be used for this region. As more structures become available, however, further canonical rules may be deduced.

CHAPTER 7 - OVERALL DISCUSSION AND FURTHER WORK

7.1 Improvements to the algorithm

An improved modelling procedure for antibody CDR-H3 loops has been presented. There are two main contributions to the improved results: first, alterations to the loop backbone building algorithm itself, and second, addition of sidechains *after* energy minimisation to give an improved backbone screen.

Alterations to the loop building

Some alterations were made in the loop backbone building itself, namely energy minimisation, ignoring 1,4-interactions, ignoring H3 sidechains in VFF and including the non-H3 CDR sidechains in CONGEN and VFF. The high energy 1,4-interactions in otherwise good structures were due to problems with the loop/framework grafting procedure and the Go and Scheraga chain closure algorithm, whereas the high energy due to the H3 sidechains appeared to be due to attempted addition of sidechains on non-minimised loops resulting in clashes between the H3 sidechains. The improvement that

inclusion of the other CDR sidechains produces shows that there must be an accurate environment when building the H3 loops, with all sidechains present. If the intra-loop 1,4 interactions and H3 sidechains are ignored but the other H3 sidechains are not included, the improvement is not seen. In this case the non-H3 sidechains were from the crystal structure; in an unknown this would not be so (but see later).

Sidechain addition on minimised structures

Sidechain addition on minimised structures, followed by screening based on accessibility patterns of the sidechains along the loop, leads to effective screening of the lowest energy conformations by VFF. The sidechain addition itself (either dead-end or CONGEN) does not effectively screen out high RMSD conformations; however the accessibility screen does, with a conformation below 2.5Å selected in 7/9 cases and 2.1Å or below in 6/9 cases. The accessibility screen effectively removes high (greater than 3.0Å) RMSD conformations in 8/9 cases. The one structure which does not do well by either measure in the accessibility screen is 1hil, which is a long loop and therefore flexible in solution, without a consistent accessibility pattern (even though its crystal structure follows the typical accessibility pattern).

In terms of sidechain prediction as measured by chi-angle accuracy, both methods, dead-end and CONGEN, predict large sidechains most frequently, and small or polar sidechains least frequently. Phe and Tyr in particular are frequently placed correctly, as misplacement would lead to clashes, whereas Ser has a large amount of conformational space available, and charged residues such as Glu and Asp can form 'false' salt-bridges if placed incorrectly. Most residue types are predicted with similar accuracy using the two methods, but Met, Tyr and Asn are predicted more accurately with dead-end, while Leu is better placed with CONGEN. The potential solution of combining the two methods is not possible, as one method would have to be employed first, and while performing that method, sidechains of the CDR to be built with the other method would have to be ignored.

The other screening method tried, namely the screen for the presence of three of the observed four framework contacts of residue 234 does not lead to improved results for all structures, either using VFF or the accessibility screen, although a noTable improvement is seen in the structure that has been repeatedly poorly modelled, namely 1hil, where the conformation with RMSD 1.7Å is selected over those with RMSD greater than 4Å. The reasons for this were discussed in Chapter 6.

The accessibility score method above appears to be more effective as a screen than the solvation energy methods of either Eisenberg and McLachlan (1986) or DelPhi (Klapper et al., 1986; Gilson and Honig, 1988). The score method takes into account the consistent pattern observed in antibody H3 loops, whereas the Eisenberg and McLachlan method assumes that hydrophobic residues will always prefer to be buried and vice-versa. This is not always true, as charged or hydrophilic residues may well be buried forming salt bridges (such as that formed in an H3 loop between residues 219 and 235 when these are Arg and Asp respectively) or hydrogen bonds, and conversely, hydrophobic residues, particularly Phe, may be exposed in order to form contacts with antigen. The problems with DelPhi are likely to relate largely to the grid representation of the molecule (Chapter 3).

7.2 Further work

Extending canonical classes to cover CDR-H3

Chapter 6 showed that it may be possible to define canonical classes for H3 loops. A number of rules have been postulated, relating largely to the C terminal region of the loop, but those defining the centre are less clear. Further

investigations need to be done on the more variable centre of the loop. We need more H3 crystal structures to attempt to elucidate sequence-structure relationships, as we do to examine whether there are other contacts formed between H3 and framework residues not as consistent as those formed by residue 234, but which may define canonical classes.

A possible approach in the absence of further H3 crystal structures is to obtain from the PDB, using the C-alpha to C-alpha distance constraints method of the AbM database search (see Introduction), loops with the same general rough shape as H3 loops. The set pulled out could then be clustered, by torsion angles or by RMSD (see Appendix 1) and each cluster examined to see whether there are consistent features of the sequence. One problem here is that framework residues determine canonical class: we would have to also record the residues within a certain distance of each database loop and direction relative to the loop.

Residue pair potentials

A second development that has been preliminarily investigated in this work is the use of residue pair potentials. There have been several attempts to calculate a knowledge-based pseudo-energy for a molecule in which the environment

of residue pairs within the molecule is classified in various ways (such as distance apart, separation on the chain, type of secondary structure, accessible surface area, and number of atoms within a given distance) and compared to the occurrence in known structures of that particular residue pair in that environment (for example, Bryant and Lawrence (1993), Abagyan et al. (1994), Willmanns and Eisenberg (1995), Godzik et al. (1992), Ouzonis et al. (1993),Gracy et al. (1993), Nishikawa and Matsuo (1993), Jones et al. (1992); for review see Fetrow and Bryant (1993)). These are usually used in threading (see Chapter 1), but in order to apply this method to screening of the CAMAL-generated conformations, it is unnecessary to apply the threading procedure, as all we need to do is evaluate the fitness of the sequence of the loop for the structure, by analysing the environment of each pair and comparing this to the frequency of that combination in known structures.

Surveying the literature has suggested that an antibody customised version of the Sippl algorithm (see Sippl (1990), Hendlich et al. (1990), Sippl and Weitckus (1992) and Flockner et al. (1995)) may be appropriate, the Sippl algorithm being relatively simple to implement but effective. Sippl's method classifies each residue-residue interaction in the protein according to the residue types, the distance

(divided into a number of equal intervals) and the sequence separation (divided into a number of classes). The c-beta atoms of the residues are the atoms used to measure the distance (c-beta atoms are added to glycines in the position that they would be in if they were 'average' alanines). To be of maximum usefulness the database of known pairs should be from antibodies, as certain interaction types occur more frequently in antibodies. This does, however, have a problem in that currently there is only a limited data set (63) of known antibody structures.

Modelling a complete Fv

Once the modelling algorithm works for the situation in which only CDR-H3 is being modelled, it needs to be tested whether it will still work if all parts of the Fv are modelled from scratch. The simplest case would be where all the other CDRs are canonical, as these would be modelled reasonably accurately, so the H3 model is unlikely to be affected to any great extent by inaccuracies in the other loops. The problem would come when more than one non-canonical loop needs to be modelled; for the first non-canonical, the other non-canonicals would have to be ignored, and for subsequent non-

canonicals, possible inaccuracies in preceding models would affect the result. The problem is compounded by the fact that sidechains of the loops other than the one being currently modelled are included in the calculations in the method as it stands. This problem needs further investigation.

Even in the cases where there are five canonicals, the problem of building the sidechains of all loops (not just H3) simultaneously needs to be addressed. Maximum overlap could be used to give an initial rough sidechain set for the canonicals for the purpose of constraining the H3 backbone conformations in the CONGEN and VFF stages, but ideally at the H3 sidechain building stage, a more accurate set should be built simultaneously for all CDRs to minimise the effect of error in the non-H3 sidechain positions on the H3 sidechain conformations.This could be done with dead-end or CONGEN. Dead-end has an advantage in that the conformations built by the CONGEN iterative algorithm are biased by the first sidechain built, whereas in dead-end this is not so: it simply attempts to find a global energy minimum. CONGEN, on the other hand, has the advantage of speed in that when considering the environment surrounding the current residue being built, a range of rotamers does not need to be considered. Alternatively the cluster search of Wilson et al (1993; see Chapter 5) which combines the advantages of

both, might be useful for building sidechains on all loops. The method has the advantage over CONGEN of not being biased to such a great extent by the first sidechain built (due to the clusters) and also includes a solvation term, unlike CONGEN. It would also be quicker than dead-end.

Finally, it may not be easy to accurately model the longer H3 loops, above 10 residues in length, as they are more flexible in solution than shorter loops and may not have a single, well-defined structure. To approach this problem, it may be necessary to resort to molecular dynamics, running a simulation of the loop (together with the rest of the Fv) in water to model its movement in solution, and as a test, compare the median of these conformations to the crystal structure.

Shirai et al (1998) have attempted a preliminary dynamics experiment, on the H3 loops of two structures: 7fab and 2fbj. Using the conformation of the other structure as a starting point, they performed dynamics on each structure and obtained quite an accurate model, particularly at the terminals. However, further studies need to be performed, firstly as a larger data set is needed, and secondly, the H3 loops of these two structures have the same length and similar conformation.

7.3 Summary of changes to the AbM algorithm

The Table below summarises the main changes to the AbM algorithm which were investigated, and shows which changes were adopted for all subsequent investigations.

Change	*Overall results/comments*
Full rather than modified VFF (Chapter 2)	No change in RMSD distribution, but adopted in order to combine with solvation-based methods subsequently (Table 2)
Steepest descent minimisation (Chapter 2)	Improvement in the RMSD spread of the bottom 200 loops. Therefore adopted. (Table 3)
Solvation energy screening by Eisenberg and McLachlan/DelPhi (Chapter 3)	No improvement, so not adopted. (Tables 7-8)
Individual VFF components as a screen (Chapter 2)	None more effective as a screen than full VFF, so not adopted. (Table 4-5)
Removal of 1,4s and H3 sidechains, and inclusion of other CDR sidechains, in CONGEN and VFF (the 'new' modelling procedure) (Chapter 5)	Improvement, so adopted. (Table 14a,14b)
Sidechain addition followed by accessibility scoring (Chapter 5)	Particularly effective screen. (Table 16)

Overall summary

The revised modelling procedure, including sidechain placement, is shown below.

Build framework and canonicals

|

Build H3 loop (CAMAL; include other CDR sidechains)

|

Strip H3 sidechains

|

VFF minimisation (no 1,4s; include other CDR sidechains)

|

Cluster bottom 200 backbones

|

Take bottom 10 and add sidechains

|

Take bottom 5 by VFF

|

Rank using accessibility score

|

Final conformation

Appendix 1. Root-mean-square deviation.

Root-mean-square deviation (RMSD) is a measure of the difference between two molecules. It is defined as:

$$RMSD = [\ \Sigma_{n=1->natoms}\ (\Delta x_n^2 + \Delta y_n^2 + \Delta z_n^2) / natoms\ ]^{1/2}$$

where Δx, Δy and Δz are the differences in x,y and z coordinates of a given atom between the two structures.

There are two measures of RMSD usually used in proteins. Firstly, there is global RMSD, which is the RMSD between two unfitted structures, and is a measure of the difference in both shape and orientation. The local RMSD between two conformations, of a segment of structure (such as a loop) by contrast, is that obtained by fitting the structures on that segment, and is therefore a measure of differences in shape between two conformations without taking into account differences in orientation. The use of local RMSD in modelling can be important as it can discriminate between incorrect conformations and those which are simply pointing in the wrong direction.

In this work, the usual measure of RMSD, unless otherwise indicated, will be global RMSD between a CDR-H3 conformation

and the crystal structure CDR-H3. Since the crystal structure conformation is being used for the framework and the other CDRs in the modelling (Chapter 2), only the RMSD between H3 atoms is measured.

Appendix 2: Clustering method.

The method used is an improvement on the method that AbM uses to cluster database loops based on torsion angles. It has be used to cluster based on either torsion angles or RMSD.

An initial stage is done in which all the conformations within a starting RMSD or torsion resolution (say 0.5Å RMSD) of the first conformation are 'clustered' with the first conformation, and all the conformations within the starting RMSD/torsion resolution of the next conformation which have not been clustered with the first conformation are clustered with that conformation, and so on.

The next stage calculates the median conformation of each cluster from the first stage and merges clusters based on whether their median conformations (as opposed to an arbitrary first conformation of the cluster, as in AbM) are within the clustering RMSD/torsion resolution (a step value, say 0.25Å, higher than the previous stage; from now on abbreviated to the *clustering resolution*). So, all clusters for which the median is within the clustering resolution of the median of the first cluster are merged with the first cluster, and then all free (i.e. not merged with the first) clusters within the clustering resolution of the median of the next free cluster are merged with it, and so on. This stage is then repeated with a higher clustering resolution until either a specified limit, or a target number of clusters is reached. The median conformations of these clusters are output.

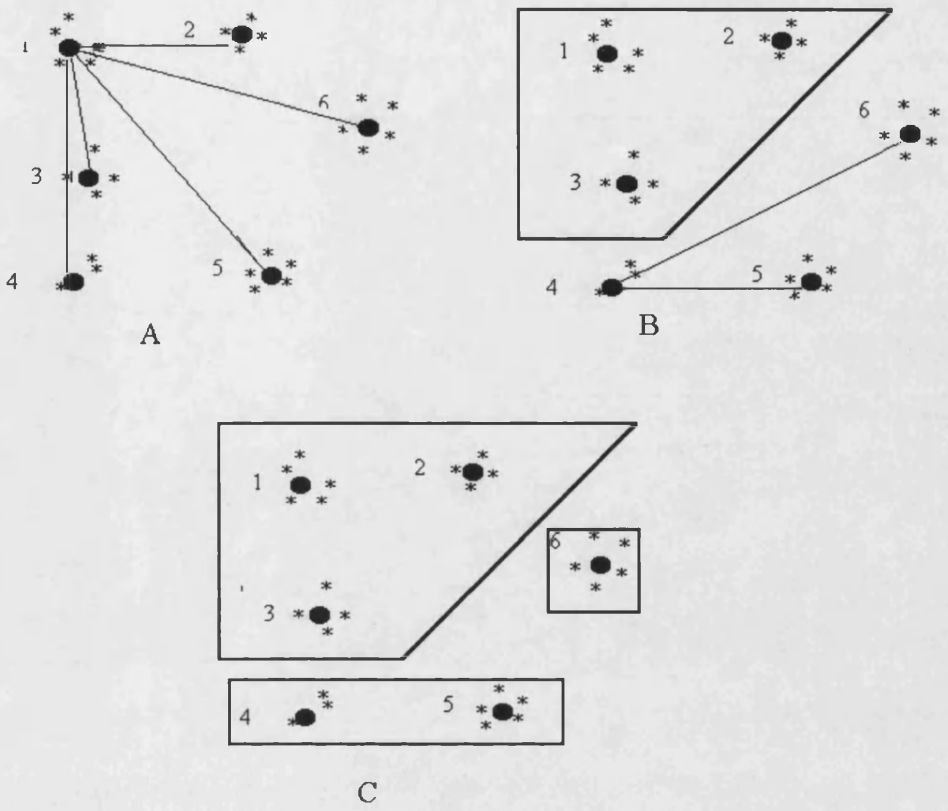


Figure 25. Median-based clustering.

The figure above illustrates the method. In **(A)** we examine which cluster medians (denoted by the dark circles) are within the clustering resolution of the median of Cluster 1. Those clusters with medians within range are clustered together with 1 (in this case Clusters 2 and 3) (B). We then move on to the next free cluster, Cluster 4, and do the same **(B),** examining only still-free clusters (i.e. 5 and 6). This results in Cluster 5 merging with 4, leaving Cluster 6 free **(C).** The resolution is then increased and the clustering process starts again. As can be seen, choosing the medians as the basis on which to merge is better than simply choosing (as in AbM) the first conformation of the

cluster come across (which could be right on the edge), but there is still some degree of arbitrariness in that the first cluster (Cluster 1 here) is chosen arbitrarily, and this leads to potential errors (for example Cluster 4 is actually closer to Cluster 3 than Cluster 5, but because Cluster 3 is merged with Cluster 1 it is out of the picture when Cluster 4 is considered). Clearly true clustering, in which every conformation is compared with every other conformation, would be ideal, but this would make unreasonable demands on time and memory. The method presented here offers a compromise.

Appendix 3. The AbM numbering system.

The residue numbering system used throughout here is the AbM system. It is a numbering system which accounts for deletions, that is, it always numbers a given residue the same in all antibodies. It has the advantage over other numbering systems, such as that of Kabat (Kabat and Wu,1992), in that it accounts for the maximum lengths observed for each loop, so that whole numbers can always be used for a residue.

L1

The L1 loop has a maximum length of 17 residues, is numbered 24-40, and is defined to lie between the conserved cysteine at position 23 and the conserved tryptophan at position 41.

L2

The L2 loop is numbered 56-62, 7 residues, and is defined to lie between the conserved tyrosine at position 55 and the conserved glycine at position 63.

L3

The L3 loop is numbered 95-105, having a maximum length of 11 residues, and is defined to lie between the conserved cysteine at position 94 and the conserved phenylalanine at position 106.

Light chain, residues 1-58 (above) 59-117 (below), illustrating the hypervariable regions.

```
                        <-------L1------>                <-L2-->
DIQMTQTTSSLSASLGDRVTISCRASQDIYN------YLNWYQQKPDGTVKLLIYYTSRLHSGV
DIVLTQSPASLSASVGETVTITCRASGNIHN------YLAWYQQKQGKSPQLLVYYTTTLADGV
EIVLTQSPAITAASLGQKVTITCSASSSVS-------SLHWYQQKSGTSPKPWIYEISKLASGV
QIVLTQSPAIMSAFPGEKVTITCSATSSVN-------YMHWFQQKPGTSPKLWIYSSSNLASGV
DVLMTQTPLSLPVSLGDQASISCRSNQTILLS-DGDTYLEWYLQKPGQSPKLLIYKVSNRFSGV
DIVMTQSPSSLTVTAGEKVTMSCTSSQSLFNSGKQKNYLTWYQQKPGQPPKVLIYWASTRESGV
DIQMTQSPSSLSASVGDRVTITCQASQDISN------YLAWYQQKPGKAPELRIYDASNLETGV

                              <---L3---->
PSRFSGSGSGTDYSLTISNLNQEDMATYICQQGNT--LPFTFGSGTKLEIKRA  1mam
PSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWS--TPRTFGGGTKLEIKR-  1vfa
PARFSGSGSGTSYSLTINTMEAEDAAIYYCQQWTY--PLITFGAGTKLELKRA  2fbj
PARFSGSGSGTSYSLTISRMEAEDAATYYCQQRSS--YPITFGSGTKLEIKRA  1for
PDRFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSH--VPPTFGGGTKLEIKRA  1igf
PDRFTGSGSGTDFTLTISSVQAEDLAVYYCQNDYS--NPLTFGGGTKLELKRA  1hil
PSRFSGSGSGTDFTFTISSLQPEDIATYYCQQYQN--LPLTFGPGTKVDIKRT  1igm
```

H1

The H1 loop is numbered 143-154, having a maximum length of 12 residues, and is defined to lie from the residue four positions after the conserved cysteine at residue 139 to the residue before the conserved tryptophan at residue 155.

H2

The H2 loop is numbered 169-180, having a maximum length of 12 residues, and is defined to lie from the residue three positions after the conserved tryptophan at residue 166 to the residue before the conserved tyrosine at residue 181.

H3

The H3 loop is numbered 220-236, having a maximum length of 17 residues, and is defined to lie from the residue three positions after the conserved cysteine at position 217 to the residue before the conserved tryptophan at position 237.

Heavy chain, residues 118-186 (above) 187-238 (below), illustrating the hypervariable regions.

```
                          <----H1---->                <----H2---->
EVKLVESGGGLVQPGGSLRLSCATSGFTFTD--YYMSWVRQPPGKALEWLGFIRNKADGYTTEYSASVK
QVQLQESGPGLVAPSQSLSITCTVSGFSLTG--YGVNWVRQPPGKGLEWLGMIW-GDG--NTDYNSALK
EVKLLESGGGLVQPGGSLKLSCAASGFDFSK--YWMSWVRQAPGKGLEWIGEIHPDSG--TINYTPSLK
QGQLQQSGAELVRPGSSVKISCKASGYAFSS--FWVNWVKQRPGQGLEWIGQIYPGDG--DNKYNGKFK
EVQLVESGGDLVKPGGSLKLSCAASGFTFSR--CAMSWVRQTPEKRLEWVAGISSGGS--YTFYPDTVK
EVQLVESGGDLVKPGGSLKLSCAASGFSFSS--YGMSWVRQTPDKRLEWVATISNGGG--YTYYPDSVK
EVHLLESGGNLVQPGGSLRLSCAASGFTFNI--FVMSWVRQAPGKGLEWVSGVFGSGG--NTDYADAVK

                                <------H3------->
GRFTISRDNSQSILYLQMNTLRAEDSATYYCTRDPYGP---------AAYWG  1mam
SRLSISKDNSKSQVFLKMNSLHTDDTARYYCARERDYR---------LDYWG  1vfa
DKFIISRDNAKNSLYLQMSKVRSEDTALYYCARLHYYGY--------NAYWG  2fbj
GKATLTADKSSTTAYMQLYSLTSEDSAVYFCARSGNYPYA-------MDYWG  1for
GRFIISRNNARNTLSLQMSSLRSEDTAIYYCTRYSSDPFY-------FDYWG  1igf
GRFTISRDNAKNTLYLQMSSLKSEDSAMYYCARRERYDENG------FAYWG  1hil
GRFTITRDNSKNTLYLQMNSLRAEDTAIYYCAKHRVSYVLTG-----FDSWG  1igm
```

Appendix 4. Solving the Poisson-Boltzmann equation by finite difference

As seen in Chapter 3, DelPhi solves the Poisson-Boltzmann equation in order to calculate the electrostatic energy of a solvated protein, taking into account the variable dielectric of the protein interior and the solvent.

The Poisson-Boltzmann equation is as follows (Nicholls and Honig, 1991):

$$\nabla [\varepsilon(x) \nabla \phi(x)] - \kappa(x)^2 \sinh(\phi(x)) = -4\pi\rho(x)$$

where $\phi(x)$ is the electrostatic potential at a particular point in space, and is determined by $\varepsilon(x)$, the spatial dielectric function, $\kappa(x)$, the Debye-Huckel term (see below) and $\rho(x)$, the charge density function.

The equation linearises to

$$\nabla [\varepsilon(x) \nabla \phi(x)] - \kappa (x)^2 \phi(x) = -4\pi\rho(x)$$

The spatially-varying dielectric function is the advantage that the method holds over simple Coulomb electrostatics, as it allows for an explicit description of the difference in dielectric constant between the protein interior and the solvent. The presence of salt is also accounted

for, by the Debye-Huckel parameter $\kappa(x)$:

$$\kappa(x) = (8\pi e^2 NI/1000kT)^{1/2}$$

where I is the bulk salt concentration (moles/litre), N is Avogadro's number, k is the Boltzmann constant, T is the temperature and e the charge on the electron.

The algorithm used to solve the Poisson-Boltzmann equation is a finite-difference method. The continuous functions (ϕ, ε, κ, ρ) are approximated by distinct values at points on a cubic grid, replacing differential operators by grid value differences. Charge is mapped to grid points making up the corners of the cube which the charge lies in, using appropriate methods to weight each corner depending on the proximity of each to the charge. The dielectric constant is set to one value inside the molecule (typically 2) and another (80) for outside, that is, in the water. 'Inside' and 'outside' are defined by solvent accessibility, i.e. any point which would at some time lie within a probe sphere of the radius of water is regarded as outside. $\kappa(x)$ is only used for grid points accessible to salt, determined by the same method as for water, using a probe sphere of the radius of a sodium ion (Klapper et al, 1986).

The grid is typically made large enough to cover the whole molecule plus a small amount extra (90% grid fill is typical) to avoid the complications at grid boundaries. The final reduction to finite

difference form (Klapper et al, 1986; Nicholls and Honig, 1991) gives:

$$\phi_o = [\ ((\Sigma_{i=1...6}\ \varepsilon_i\ \phi_i) + 4\pi\ q_o/h\ ) / ((\Sigma_{i=1...6}\ \varepsilon_i) + (\kappa_o\ h)^2\ ]$$

where ϕ_o is the potential at a particular grid point, ϕ_i the potential at the six nearest neighbours, εi the dielectric constant at the midpoint between ϕ_o and ϕ_i, qo the charge assigned to the grid point and h is the grid spacing in angstroms.

The electrostatic energy of the entire molecule is then obtained by summing the product of the potential and charge at each grid point.

References

Abagyan R, Frishman D and Argos P, Proteins 1994,**19**,p132

Alexander J, Protoplasma 1931, **14**, p296

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ, J.Mol.Biol. 1990, **215**, p403

Bairoch A, Nucleic Acids Res. 1991, **19**, p2241

Barry MM, Mol CD, Anderson VF and Lee JS, J.Biol.Chem. 1994, **269**, p3623

Bernstein F, Koetzle T, Williams G, Meyer E, Brice M, Rodgers J, Kennard O, Shimanouchi T and Tasumi M, J.Mol.Biol. 1977, **112**, p535

Bhat TN, Bentley TA, Boulot G, Green MI, Tello D, Dall'Acqua W, Souchon H, Schwarz FP, Mariuzza RA and Poljak RJ, Proc. Natl. Acad. Sci. USA 1994, **91**, p1089

Blundell TL, Carney D, Gardner S, Hayes F, Howlin B, Hubbard T et al., Eur J.Biochem. 1988, **172**, p513

Blundell TL, Sibanda BL, Sternberg MJE and Thornton JM, Nature 1987, **326**, p347

Born M, Z.Physik 1920, **1**, p45

Brady RL, Edwards DJ, Hubbard RE, Jiang JS, Lange S, Roberts SM et al., J.Mol.Biol. 1992, **227**, p253

Breinl F and Haurowitz F, Z.Physiol.Chem. 1930, **192**, p45

Brooks CL III, Karplus M, Pettitt BM, in "Proteins - A Theoretical Perspective of Dynamics, Structure and Thermodynamics", 1988.

Bruccoleri RE, Haber E and Novotny J, Nature 1988, **335**, p564

Bruccoleri RE and Karplus M, Biopolymers 1987, **26**, p137

Bryant SH and Lawrence CE, Proteins 1993, **16**, p92

Bundle DR, Baumann H, Brisson JR, Gagne SM, Zdanov A and Cygler M, Biochemistry 1994, **33**, p5183

Burkert U and Allinger NL, in "Molecular Mechanics", 1982.

Bystroff C and Baker D, Proteins 1997, **S1**, p167

Chothia C and Lesk AM, J.Mol.Biol. 1987, **196**, p901

Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, et al., Nature 1989, **342**, p877

Chou PY and Fasman GD, Biochemistry 1974, **13**, p222

Cohen FE, Abarbanel RM, Kuntz ID and Fletterick RJ, Science 1986, **234**, p349

Cohen FE, Novotny J, Sternberg MJE, Campbell DG and Williams AF, Biochem.J. 1981, **195**, p31

Cohen FE, Richmond TJ and Richards FM, J.Mol.Biol. 1979, **132**, p275

Cohen FR and Sternberg MJE, J.Mol.Biol. 1980, **137**, p9

Cohen FE, Sternberg MJE and Taylor WR, Nature 1980, **285**, p378

Constantine KL, Friedrichs MS, Goldfarb V, Jeffrey PD, Sheriff S and Mueller L, Proteins 1993, **15**, p290

Curtis BM, Presnell SR, Srinivasan S, Sassenfeld H, Klinke R, Jeffrey E, et al., Proteins 1991, **11**, p111

Dauber-Osguthorpe P, Roberts VA, Osguthorpe DJ, Wolff J, Genest M and Hagler AT, Proteins 1988, **4**, p31

Dayhoff MO, Schwartz RM and Orcutt BC, in "Atlas Of Protein Sequence and Structure" chap.5 , p345, 1978.

de la Paz P, Sutton BJ, Darsley MJ and Rees AR, EMBO J., 1986, **5(2)**, p415

Desmet J, De Maeyer M, Hazes B and Lasters I, Nature 1992, **356**, p539

Devereux J, Haeberli P and Smithies O, Nucleic Acids Res. 1984, **12**, p387

Dreyer WJ and Bennett JC, Proc.Natl.Acad.Sci.USA 1965, **54**, p864
Ehrlich P, Klin.Jahrb. 1897, 60, p299 (translated in "The Collected Papers of Paul Ehrlich", vol.2 p.107, 1957).

Eigenbrot C, Randal M, Presta L, Carter P and Kossiakoff AA, J.Mol.Biol. 1993, **229**, p969

Eisenberg D and McLachlan AD, Nature 1986, **319**, p199

Evans JV, Rose DR, To R, Young NM and Bundle DR, J. Mol. Biol. 1994, **241**, p691

Fan Z, Shan L, Guddat LW, He X, Gray WR, Raison RL and Edmundson AB, J. Mol. Biol. 1992, **228**, p188

Fauchere JL and Pliska V, Eur.J,.Med.Chem.Chim.Ther., 1983, **18**, p369

Feng DF, Johnson MS and Doolittle RF, J.Mol.Evol. 1985, **21**, p112

Fetrow JS and Bryant SH, Biotechnology 1993, **11**, p479

Fine RM, Wang H and Shenkin PS, Yarmush DL and Levinthal C, Proteins 1986, **1**, p342

Fitch WM, J. Mol. Biol. 1966, **16**, p9

Floeckner H, Braxenthaler M, Lackner P, Jaritz M, Ortner M and Sippl MJ, Proteins 1995, **23**, p376

Foote J and Milstein C, Proc. Nat. Acad. Sci. USA 1994, **91**, p10370

Foote J and Winter G, J.Mol.Biol. 1992, **224**, p487

Frishman D and Argos P, Prot.Eng. 1996, **9**, p133

Fuchs R, CABIOS 1990, **7**, p257

Garnier J, Osguthorpe DJ and Robson B, J.Mol.Biol. 1978, **120**, p97

Garnier J, Gibrat JF and Robson B, Meth. Enz, 1996, **266**, p540

Gelin B and Karplus M, Nature 1975, **272**, p222

Gilson MK and Honig B, Proteins 1988, **4**, p7

Go N and Scheraga HA, Macromolecules 1970, **3**, p178

Godzik A, Kolinski A and Skolnik J, J.Mol.Biol. 1992, **227**, p227

Goldman N, Thorne JL and Jones DT, J. Mol. Biol. 1996, **263**, p196

Goldstein RF, Biophys.J., 1994, **66**, p1335

Gracy J, Chiche L and Sallantin J, Prot. Eng. 1993, **6**, p821

Gregoret LM and Cohen FE, J.Mol.Biol. 1990, **211**, p959

Gregory DS, Martin ACR, Cheetham JC and Rees AR, Prot. Eng. 1993, **6**, p29

Guddat LW, Shan L, Anchin JM, Linthicum DS and Edmundson AB, J. Mol. Biol. 1994, **236**, p247

Hagler A, Huber E and Lifson S, J.Am.Chem.Soc. 1974, **96**, p5319

Haurowitz F, Biol Rev Camb Phil Soc, 1952, **27**, p247

Havel TD and Snow ME, J.Mol.Biol. 1991, **217**, p1

Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ, and Cobleigh T, J.Mol.Biol. 1990, **216**, p167

Hirschfelder JO, Curtiss CF and Bird RB, in "Molecular Theory of Gases and Liquids", 1964.

Holm L and Sander C, J.Mol.Biol. 1991, **218**, p183

Holmes MA and Foote J, J. Immunol.. 1997, **158**, p2192

Hsieh-Wilson LC, Schultz PG and Stevens RC, Proc. Natl. Acad. Sci. USA 1996, **93**., p5363

Jones DT, Proteins 1997, **S1**, p191

Jones DT, Taylor WR and Thornton JM, Comput. Appl. Biosci. 1992, **8**, p275

Jones DT, Taylor WR and Thornton JM, Nature 1992, **358**, p86

Kabat E and Wu T, in "Sequences of proteins of immunological interest - 5th edition", 1992.

Kabsch W and Sander C, Biopolymers 1983, **22**, p2577

Karlin S and Ghandow G, Proc.Natl.Acad.Sci.USA 1985, **82**, p6186

King RD and Sternberg MJE, J.Mol.Biol. 1990, **216**, p441

Klapper I, Hagstrom R, Fine R, Sharp K and Honig B, Proteins 1986, **1**, p47

Kurochkina N and Lee B, Prot. Eng. 1995, **8**, p437

Landau LD and Lifshitz EM, in "Quantum Mechanics" 3rd edition sect.89, 1977.

Laskowski RA, Macarthur MW, Moss DS and Thornton JM, J.Appl.Cryst. 1993, **26 Pt 2**, p283

Lasters I, De Maeyer M and Desmet J, Prot. Eng. 1995, **8**, p815
Laughton CA, J.Mol.Biol. 1994, **235**, p1088

Lee C and Subbiah S, J.Mol.Biol. 1991, **217**, p373

Levin JM, Pascarella S, Argos P and Garnier J, Prot. Eng. 1993, **6**, p849

Levitt M, J.Mol.Biol. 1992, **226**, p507

Lim VI, J.Mol.Biol. 1974, **88**, p873

Liu H, Smith TJ, Lee W, Mossert G, Rueckert BR, Olson NH et al., J.Mol.Biol. 1994, **240**, p127

Mandal C, Kingery BD, Anchin JM, Subramaniam S and Linthicum DS, Nature Biotechnology 1996, **14**, p323

Martin ACR, Cheetham JC and Rees AR, Meth.Enz. 1991, **203**, p121

Martin ACR, Cheetham JC and Rees AR, Proc. Natl. Acad. Sci. USA 1989, **86**, p9268

McCammon JA and Harvey SC, "Dynamics of Proteins and Nucleic Acids", 1987.

McLachlan AD, J. Mol. Biol. 1972, **64**, p417

Morea V, Tramontano A, Rustici M, Chothia C and Lesk AM, J.Mol.Biol. 1998, **275**, p269

Mudd J, J.Immunol. 1932, **23**, p243

Muggleton S, King RD and Sternberg MJE, Prot. Eng. 1992, **5**, p647.

Needleman SB and Wunsch CD, J.Mol.Biol. 1970, **48**, p443

Nicholls A and Honig B, J. Comp. Chem. 1991, **12(4),** p435

Nishikawa K and Matsuo Y, Prot. Eng. 1993, **6(8),** p811

Northrup SH, Pear MR, Morgan JD, McCammon JA and Karplus M, J.Mol.Biol. 1981, **153**, p1087

Oliva B, Bates PA, Querlo E, Aviles FX and Sternberg MJE, J. Mol.Biol. 1997, **266**, p814

Ouzonis C, Sander C, Scharf M and Schneider R, J.Mol.Biol. 1993, **232**, p805

Padlan EA, Davies DR, Recht I, Guval D and Wight C, Cold Spring Harbour Symp. Quant. Biol., 1976, **41**, p627

Pauling L, J.Am.Chem.Soc. 1940, **62**, p2643

Pearson WR and Lipman DJ, Proc.Natl.Acad.Sci.USA 1988, **85**, p2444

Pedersen J, Searle S, Henry AH and Rees AR, Immunomethods 1992, **1**, p126

Pedersen JT, Henry AH, Searle SJ, Guild BC, Roguska M and Rees AR, J.Mol.Biol. 1994, **235**, p939

Pedersen JT and Moult J, Proteins 1997, **S1**, p179

Podlogar BL, Leo GC, McDannell PA, Loughney DA, Caldwell GW and Barrett JF, J.Med.Chem 1997, **40**, p3453

Posfai J, Bhagwat AS, Posfai G and Roberts RJ, Nucleic Acids Res 1989, **17**, p2421

Presnell SR, Cohen BI and Cohen FE, Comput. Appl. Biosci. 1993, **9**, p373

Pulito VL, Roberts VA, Adair JR, Rothermel AL, Collins AM, Varga SS, Martocello C, Bodmer M, Jolliffe LK and Zivin RA, J. Immunol 1996, **156**, p2840

Qian N and Sejnowski TJ, J. Mol. Biol. 1988, **202**, p865.

Rees AR, Searle SMJ, Henry AH and Pedersen JT, in "Protein Structure Prediction" (edited by Sternberg MJE), 1996.

Reichman L, Clark M, Waldmann H and Winter G, Nature 1988, **332**, p323

Richmond TJ and Richards FM, J.Mol.Biol. 1978, **119**, p537

Rice DW and Eisenberg D, J.Mol.Biol. 1997, **267**, p1026

Roguska MA, Pedersen JT, Keddy CA, Henry AH, Searle SJ, Lambert JM, Gottmacher VS, Blattler WA, Rees AR and Guild BC, Proc. Natl. Acad. Sci. USA 1994, **91**, p969

Roguska MA, Pedersen JT, Searle SMJ, Henry AH, Rees AR et al, Prot. Eng. 1996, **9**, p895

Rose G, Geselowitz A, Lesser G, Lee R and Zehfus M, Science 1985, **229**, p834

Rost B and Sander C, Proteins 1995, **23**, p295

Rost B, Schneider R and Sander C, J.Mol.Biol. 1997, **270**, p471

Russell RB, Saqi MAS, Sayle RA, Bates PA and Sternberg MJE, J.Mol.Biol. 1997, **269**, p423

Russell RB, Saqi MAS, Bates PA, Sayle RA and Sternberg MJE, Prot. Eng. 1998, **11**, p1

Salamov AA and Solovyev VV, J.Mol.Biol. 1995, **247**, p11

Sali A and Blundell TL, J.Mol.Biol. 1993, **234**, p779

Samudrala R and Moult J, J.Mol.Biol. 1998, **279**, p287

Scheraga HE, Adv Phys Org Chem 1968, **6**, p103

Schulze-Gahnen U, Rini JM and Wilson IA, J. Mol. Biol. 1993, **234**, p1098

Sheriff S, Chang CY, Jeffrey PD and Bajorath J, J.Mol.Biol. 1996, **259**, p938

Shirai H, Kidera A and Nakamura H, FEBS Letters 1996, **399**, p1

Shirai H, Nakajima N, Higo J, Kidera A and Nakamura H, J.Mol.Biol. 1998, **278**, p481

Sibanda BL and Thornton JM, J.Mol.Biol. 1993, **229**, p428

Sippl MJ and Weitckus S, Proteins 1992, **13**, p258

Sippl MJ, J.Mol.Biol. 1990, **213**, p859

Sitkoff D, Sharp KA and Honig B, J.Phys.Chem. 1994, **98**, p1978

Smith HO, Annau TM and Chandrasegaran S,
Proc.Natl.Acad.Sci.USA 1990, **87**, p826

Smith KC and Honig B, Proteins 1994, **18**, p119

Smith TF and Waterman MS, J.Mol.Biol. 1981, 147, p195

Stanfield RL, Fieser TM, Lerner RA and Wilson IA, Science 1990,
248, p712

Stanford JM and Wu TT, J. Theor. Biol., 1981, **88**, p481

Sternberg MJE, in "Protein Structure Prediction" (edited by Sternberg
MJE), 1996.

Suh SW, Bhat TN, Novia MA, Cohen GH, Rao DN, Rudikoff S, and
Davies DR, Proteins 1986, **1**, p74

Teller D, Nature 1976, **260**, p729

Tonegawa S, Nature 1983, **302**, p575

Warshel A and Lifson S, J.Chem.Phys. 1970, **53**, p582

Weiner SJ, Kollman PA, Case DA, Singh UC, Ghia C, Alagona G and
Weiner P, J.Am.Chem.Soc. 1984, **106**, p765

Wilmanns M and Eisenberg D, Prot.Eng. 1995, **8(7)**, p627

Wilson C, Gregoret LM and Agard DA, J.Mol.Biol. 1993, **229**, p996

Wilson EB, Decius JC, and Cross PC, in "Molecular Vibrations"
1955.

Zhu ZY and Blundell TL, J. Mol. Biol. 1996, **260**, p261